HHS Public Access

Author manuscript

*Nature*. Author manuscript; available in PMC 2016 April 08.

Published in final edited form as:

*Nature*. 2015 October 8; 526(7572): 253–257. doi:10.1038/nature15390.

# A novel locus of resistance to severe malaria in a region of ancient balancing selection

**Malaria Genomic Epidemiology Network 1**

The high prevalence of sickle haemoglobin in Africa shows that malaria has been a major force for human evolutionary selection, but surprisingly few other polymorphisms have been proven to confer resistance to malaria in large epidemiological studies[1–3]. To address this problem we conducted a multi-centre genome-wide association study (GWAS) of life-threatening *Plasmodium falciparum* infection (severe malaria) in over 11,000 African children with replication data in a further 14,000 individuals. Here we report a novel malaria resistance locus close to a cluster of genes encoding glycophorins that are receptors for erythrocyte invasion by *P. falciparum*. We identify a haplotype at this locus which provides 33% protection against severe malaria (OR=0.67, 95%CI=0.60–0.76, $P=9.5\times10^{-11}$) and is linked to polymorphisms previously shown to have features of ancient balancing selection, based on haplotype sharing between humans and chimpanzees[4]. Taken together with previous observations on the malaria-protective role of blood group O[1–3,5], these data reveal that two of the strongest GWAS signals for severe malaria lie in or close to genes encoding the glycosylated surface coat of the erythrocyte cell membrane, both within regions of the genome where it appears that evolution has maintained diversity for millions of years. These findings provide new insights into the host-parasite interactions that are critical in determining the outcome of malaria infection.

In the discovery phase of this study we analysed GWAS data on 5,633 children with severe malaria and 5,919 population controls from The Gambia, Kenya and Malawi, and in the replication phase we analysed candidate SNPs in a further 13,946 case control samples from Burkina Faso, Cameroon, The Gambia, Ghana, Malawi, Mali and Tanzania (Extended Data Figure 1). The majority of samples used in the discovery phase have been analysed previously by lower resolution GWAS methods[3]. For this analysis we improved resolution by directly typing all samples at approximately 2.5 million SNPs using the Illumina Omni2.5M platform, followed by quality control (Extended Data Figures 1 and 2) and imputation of genotypes at over 10 million SNPs using haplotype data from the 1000



Correspondence and requests for materials should be addressed to Dominic Kwiatkowski (dominic@sanger.ac.uk) or Chris Spencer (chris.spencer@well.ox.ac.uk).

[1]A list of authors is included at the end of the manuscript.

**Supplementary Information** is linked to the online version of the paper at www.nature.com/nature.

**Author Information** Genotype and phenotype data underlying this manuscript will be deposited in the European Genome-phenome Archive (accession EGAS00001001311). Access to individual-level genotype data is available by application to an Independent Data Access Committee: see www.malariagen.net/data. For further details of data underlying this manuscript, see http://www.malariagen.net/resource/14.

The authors declare no competing financial interests.

Genomes Project[6]. Imputation performance varied across populations, with accuracy decreasing as a function of the similarity between study and reference individuals (Extended Data Figure 3). When testing for genetic association, principal components analysis was used to correct for population structure (Extended Data Figure 4a–e) which reflected both geography and self-reported ethnicity. Similar results were obtained using a mixed-model approach (Extended Data Figure 4f).

To assess the evidence for association in the discovery phase we used an approach which allows for heterogeneity in the protective effect of an allele across different study sites. This could be particularly important in our data as high levels of genetic and ethnic diversity in Africa can result in variable patterns of linkage disequilibrium between study sites that can complicate GWAS analysis[7]. Other potential sources of heterogeneity include allelic heterogeneity and multiple independent origins of malaria resistance loci, as has been well documented at the *HBB* locus[1,3], as well as the high levels of genetic diversity in the parasite[8]. Specifically, we used a Bayesian approach that combines evidence across multiple models of association by specifying a prior probability on the size and similarity of genetic effect across populations, as well as the mode of inheritance[1]. A single statistical summary of the signal of association was obtained by averaging the evidence across models, weighting each by its prior probability, and comparing the evidence to the null model of no association (model averaged Bayes factor; $BF_{avg}$). Having observed the data, a posterior probability was assigned to each model ($Pr_{\text{model}}$) conditional on it being a true association and the model assumptions, which are described in the Methods and Extended Data Figure 5. We replicated previously reported GWAS signals[2,3,9] at the *HBB* ($BF_{avg}=5.8\times10^{24}$), *ABO* ($BF_{avg}=6.7\times10^{9}$) and *ATP2B4* ($BF_{avg}=4.4\times10^{5}$) loci, and a detailed analysis of key variants at these and other previously reported loci is presented elsewhere[1]. A previously reported association near the gene *MARVELD3*[2] did not replicate in this dataset (Supplementary Note 1). Genome-wide patterns of association with severe malaria at the 34 regions of the genome containing a variant with either a Bayes factor for the most probable model ($BF_{max}$) > $2.5\times10^{4}$ or with a model averaged Bayes factor ($BF_{avg}$) > $2.5\times10^{3}$ are summarised in Extended Data Figure 6 and Supplementary Table 1. Details of the evidence for association in these regions can be viewed online at www.malariagen.net/resource/14.

These data provide a rich resource of new candidate loci for further investigation. Here we focus on a region of chromosome 4 shown in Figure 1, where the strongest signal of association (at SNP rs184895969) is located between the gene *FREM3* and a cluster of three glycophorin genes (*GYPE, GYPB* and *GYPA*). Glycophorins are sialoglycoproteins that are abundantly expressed in the erythrocyte membrane, providing a hydrophilic surface coat that is necessary for erythrocytes to flow freely in the circulation. A complex system of single-nucleotide and structural variants in this region determine the MNS blood group system[10]. These genes play a functional role in invasion of erythrocytes by *P. falciparum*. Glycophorin A is the receptor for the *P. falciparum* erythrocyte-binding ligand EBA-175[11], and glycophorin B is a receptor for the parasite ligand EBL-1[12]. To follow up this observation, selected SNPs at this locus were genotyped by Sequenom iPLEX MassArray in the discovery and replication sample sets outlined above (Figure 1 and Extended Data Figure 7a). The combined dataset of 25,498 samples provided convincing evidence of association at

rs186873296 by standard fixed-effect meta-analysis ($P = 9.5\times10^{-11}$) as well as by the above Bayesian approach ($BF_{overall} = 1.3\times10^{8}$; Figure 2 and Methods). The minor allele frequency of rs186873296 was higher in East Africa than West Africa, and the greatest evidence of association was seen in Kenya where it was most common, with an allele frequency of approximately 10%. Using only replication data to avoid winner's curse, and assuming an additive model, we estimate that carrying one copy of the derived (non-ancestral) allele reduces the risk of severe malaria by about 40% in Kenya (OR = 0.60, 95% CI=0.46–0.79), with a slightly smaller effect across all populations (OR = 0.67, 95%CI 0.56–0.80 in frequentist fixed-effect meta-analysis). Further details are given in Supplementary Note 2.

The glycophorin gene cluster has a complex pattern of gene conversion and structural variation that has been previously noted; indeed it has been proposed that selective pressure due to pathogens, including malaria, has contributed to shaping diversity in this region[10,13–16]. Using the human reference sequence and mapped sequence read data from the 1000 Genomes Project, we identified the boundaries of a 350kb region of sequence homology surrounding these genes as well as a set of segregating gene deletions (Extended Data Figure 8). The lead imputed marker (rs184895969) is located within 10kb of this complex region. Imputation accuracy within the region is low using current reference data, so it is possible that the causal variant lies within the glycophorin genes themselves but that this is obscured in the current imputed dataset. With this caveat, we computed a credible set of putatively causal variants in the region; this set includes both the lead imputed marker and a linked missense mutation (rs181620317) in *FREM3* (Extended Data Figure 7b). We note that the protective allele at rs184895969 was associated with increased *GYPA* transcription in published gene expression data for HapMap lymphoblastoid cell lines[17] ($P = 0.016$; Extended Data Figure 9); other regional analyses are described in Supplementary Notes 1–2. Improved African genome variation reference panels[7,18] are needed to understand the complex patterns of variation in this region so that the causal variant of malaria resistance can be fine-mapped with greater confidence.

A striking feature of these data is that all of the loci which reach conventional criteria for genome-wide significance ($P < 5\times10^{-8}$ using a fixed effects model) are in or near genes that play a key role in erythrocytes (*HBB, ABO, ATP2B4, FREM3/GYPE*), the primary host cell of *P. falciparum*. Other erythrocyte-related genes are identified in the discovery phase analysis but do not reach genome-wide significance, including *EPB41*, which encodes erythrocyte membrane protein band 4.1 and has been implicated as a possible receptor for *P. falciparum* invasion[19] (rs2985337: $BF_{avg}$=3443; fixed-effect meta-analysis *OR*=1.16, 95%CI=1.09–1.23, $P=1.2\times10^{-6}$; see Extended Data Figure 6).

An intriguing feature of the *ABO* locus is that it contains a number of polymorphisms that are shared between humans and other primates, and recent analyses of sequence variation across species indicated that some of these are ancient polymorphisms that have been maintained by balancing selection over millions of years[20]. The current findings are of particular interest since the *FREM3/GYPE* is one of the most prominent examples of putative ancient balancing selection in a genome-wide analysis of haplotype sharing between humans and chimpanzees[4]. The peak GWAS signal at this locus is less than 45kb away from the shared human-chimp haplotype, which falls within the region covered by the

credible set (Extended Data Figure 7b), although it does not exhibit a strong association with severe malaria (Supplementary Note 3). To explore the genealogical relationship between the putative ancient balanced polymorphisms (ABPs) and SNPs associated with severe malaria in the *FREM3/GYPE* region, we inferred an ancestral tree[21] from the African (YRI + LWK) part of the 1000 Genomes data and used it to order haplotypes in the region, labelled with the positions of ABPs, malaria-associated SNPs, and other variants of interest (Figure 3). All three haplotypes carrying the protective allele at the directly-typed marker with most evidence of association (rs186873296) carry the minor allele at the ABP markers (D'=1, P=0.017). By inferring the positions of putative causal mutations on the estimated genealogical tree[21] at the lead imputed marker (rs184895969) we found evidence for a single protective mutation in Kenya (*log10 BF*=3.09; *OR*=0.6) estimated to lie on the branch ancestral to the protective allele at the lead marker. Although the most likely position for an additional mutation was on the branch ancestral to the ABPs, a single haplotype explains most of the signal of association in this region.

These observations raise the question of whether malaria resistance loci are more likely to be found in regions of the genome that show evidence of ancient balancing selection. We therefore analysed the relationship between the regions of association in our GWAS and 125 regions of the genome found by Leffler et al[4] to contain haplotypes shared between humans and chimpanzees. The SNPs defining these haplotypes (ABPs) were not themselves enriched for association with severe malaria ($P > 0.1$, Methods). We used a simulation approach to assess the physical proximity of ABPs to the peak of association within the 34 strongest regions of association (tier 1) and 73 weaker signals (tier 2), and observed a significant relationship with tier 1 over a range of length scales (Extended Data Figure 10a,d). We also identified the nearest gene to the lead marker within each association region as well as the gene nearest to each ABP haplotype, and performed a gene-based test for genome-wide enrichment. Strong evidence for enrichment was seen at tier 1 loci (OR 41.0; $P=4\times10^{-7}$ by Fisher's exact test, $P_{sim}=5\times10^{-4}$ using a simulation approach described in Methods, and $P=1\times10^{-4}$ using the INRICH algorithm[22]; Extended Data Figures 10b–d) and a weaker trend at tier 2 loci (OR 7.7, $P_{sim}=0.15$). Apart from *ABO* and *FREM3/GYPE*, there were six other GWAS loci (4 in tier 1, 2 in tier 2) where the nearest gene to the lead marker was also the nearest gene to an ABP haplotype (*DSCAM, NRG1, CNTNAP5, TMEM132C, CACNA2D1, RYR2*). Although the current association evidence at these loci do not satisfy conventional criteria for genome-wide significance and they should be regarded as putative until convincingly replicated, it is noteworthy that they are all involved in key aspects of membrane biology (Supplementary Note 3).

In the largest genetic association study of malaria to date, we have identified a new locus of resistance to severe malaria that lies next to a cluster of glycophorin genes involved in erythrocyte invasion by *P. falciparum*, and that also overlaps a locus of putative ancient balancing selection identified by analysis of haplotype sharing between humans and chimpanzees. It is possible that malaria is not the cause of the ancient balancing selection, or that it is just one of a number of opposing evolutionary driving forces, as at *ABO*, where blood group O reduces the risk of severe malaria but increases the risk of severe cholera[23]. Nonetheless, these new findings raise the intriguing question of whether natural selection on

malaria susceptibility has been shaping genetic diversity in humans and their ancestors for millions of years. *P. falciparum* is closely related to the chimpanzee parasite *P. reichenowi* and other parasites of African Great Apes[24–26]. It has been proposed that *P. falciparum* was introduced into human populations from chimpanzees or gorillas in the recent past, but this remains a matter of intense debate[25–27]. Population genetic data are consistent with an ancient origin followed by a marked expansion of the parasite population approximately ten thousand years ago, coincident with the introduction of agriculture[28]. The *P. falciparum* genome possesses a huge repertoire of polymorphism[8] and it is possible that the host and parasite genomes are engaged in a longstanding evolutionary arms race, each maintaining diversity to try to outflank the other[29]. Intriguingly, the parasite surface receptor EBA-175, which directly binds glycophorin A during red cell invasion, also contains structural polymorphisms with features of ancient dimorphism[30]. The present findings provide new leads both to investigate these evolutionary mechanisms and to discover further genetic determinants of resistance to severe malaria in African children.

## Methods

### Samples, ethics and clinical information

Samples were collected from nine partner projects from across sub-Saharan Africa (Extended Data Figure 1a) as described previously[1,3]. The studies and sample sets described in this manuscript form part of a larger ongoing project within the Malaria Genomic Epidemiology Network (www.malariagen.net). We used the World Health Organisation definition of severe malaria which comprises a broad spectrum of life-threatening clinical complications of *Plasmodium falciparum* infection[29,31]. Investigators from study sites worked together to agree on principles for sharing data and standardising clinical definitions, and to define best ethical practices across different local settings including the development of guidelines for informed consent. Relevant ethics committees are listed in Extended Data Figure 1a. Further information on policies, research and the consent process may be found on the MalariaGEN website (http://www.malariagen.net/community/ethics-governance).

### DNA extraction and Sequenom typing

As described previously[1,3], all samples submitted to the MalariaGEN Resource Centre underwent a standard set of procedures that included quantification using picogreen, genotyping of 65 polymorphisms (including HbS - rs334, and 3 gender-typing SNPs) on the Sequenom iPLEX MassArray platform and matching to baseline clinical data (e.g. gender, ethnic group and case-control status).

### High-density genotyping

Three cohorts (Gambia, Malawi and Kenya) were genotyped on the Illumina HumanOmni2.5–4 (Kenya) and Illumina HumanOmni2.5–8 (Gambia, Malawi) platforms. As described previously[3] we used three different calling algorithms (Illuminus[32], Illumina's Gencall algorithm as provided in BeadStudio, and GenoSNP[33]), each of which uses slightly different information in the data. We formed final genotype calls by taking consensus between the three algorithms. Genotypes where any two of the three calling algorithms were discordant, and genotypes where fewer than two algorithms were confident enough to make

a call were treated as missing. This process showed improved calling, evaluated using Mendelian error counts in a subset of Kenyan samples, relative to each of the three algorithms separately (data not shown).

Following genotype calling, we aligned genotypes to the forward strand of the human reference sequence (GRCh37), using both the Illumina-supplied manifest and publically available strand files (www.well.ox.ac.uk/~wrayner/strand) obtained by mapping allele probes to the reference by BLAT[34]. We removed SNPs whose position or strand mismatched between the Illumina manifest and the strand file. To simplify analyses, we restricted attention to the set of SNPs having the same name, chromosome, position, strand, and probe sequences across the two genotyping platforms. Because the Omni2.5M contains multiple probes for some variants, we further removed SNPs to ensure positions were unique. In total we were left with 2322985 SNPs in each cohort across the autosomes and the X and Y chromosomes. We note that SNPs annotated as lying on the pseudo-autosomal region (PAR) of the X chromosome were not included, as these had position equal to 0 in the manifest for the HumanOmni2.5–4 array. Finally, we flipped alleles where necessary so that in downstream analyses the first allele always corresponds to the reference allele of the human genome sequence.

## Sample Quality Control

We performed sample QC separately on each cohort by computing autosome-wide averages of normalised X channel intensity, normalised Y channel intensity, and heterozygosity and missingness based on the consensus call. To identify outlying samples we applied ABERRANT[35], adjusting the lambda parameter per cohort to account for differences in genotyping quality (Extended Data Figure 1d–f). In Gambia, a tail of samples showing low heterozygosity but otherwise appearing to be well typed was apparent. We explicitly included these samples in downstream analyses.

To estimate genome-wide relatedness between samples, we selected a list of 178775 high quality SNPs satisfying the criteria missingness < 1%, MAF > 1% and thinned to be at least 0.005cM apart and to have pairwise $r^2 < 0.3$. Treating each cohort separately, we used SHELLFISH (www.stats.ox.ac.uk/~davison/software/shellfish/shellfish.php) to compute a matrix of pairwise relatedness values $R = (r_{ij})$ between samples, where $r_{ij}$ denotes the genome-wide average covariance of frequency-normalised genotypes in individuals i and j. In samples with few close relationships, the value of $r_{ij}$ can be thought of as an estimate of kinship[36] with, for example, values close to 1 representing identity between samples, and values close to zero reflecting a lack of close relatedness (relative to the rest of the sample). To remove samples with duplicate typing and close relationships, we excluded one of each pair of samples with $r_{ij} > 0.2$, taking all remaining samples through to phasing and imputation. Extended Data Figure 1c lists the number of samples before and after QC, and the number removed by each QC step; (*) denotes the number of samples removed after explicitly including samples with low heterozygosity in Gambia.

We used SHELLFISH to compute principal components (PCs) on the post-QC sample set. Consistent with our removal of poor quality, duplicate and closely related samples, principal components plots show no substantially outlying samples (see Extended Data Figure 4a–c).

The top few principal components in each cohort reflect substantial population structure, as evidenced by colouring by reported ethnicity. As found previously[3,9], the top principal components are also significantly correlated with case-control status (Extended Data Figure 4d), indicating that population structure may act as a confounding factor in association analyses if not controlled.

## Genotyping Quality Control

Treating each cohort separately, we used SNPTEST (mathgen.stats.ox.ac.uk/genetics_software/snptest/snptest.html) to test for association at each autosomal SNP using a genotypic model of association which allows for different effects at heterozygote and homozygote genotypes, including the leading 5 principal components as covariates. Genotypic model association tests are particularly sensitive to confounding by genotyping error[37], and we used the resulting *P*-values as a guide to finding appropriate QC criteria. To detect potential spurious genotypes due to batch effects, we modelled genotypes as predicted by the leading 5 principal components and case/control status in a linear regression framework in R[38], and tested whether including an indicator of the plate on which each sample was genotyped contributed significantly to model fit. We refer to this as the "Plate test". For downstream analyses we excluded SNPs with minor allele frequency < 1%, missing data proportion > 5% (in Gambia and Malawi) or > 2.5% (in Kenya, which had fewer missing genotypes overall), Hardy-Weinberg $P < 1\times10^{-20}$ in controls, or Plate-test $P < 0.01$. We inspected cluster plots of all remaining SNPs showing association test $P < 1\times10^{-5}$ (Gambia) or $P < 1\times10^{-4}$ (Kenya, Malawi) and excluded those with clear genotyping problems. Extended Data Figure 2a shows the number of SNPs excluded by each criterion. The post-exclusion genome-wide association analysis is shown in Extended Data Figure 2c.

SNP QC on the X chromosome was performed as on the autosomes, with a few differences as follows. We treated males and females separately, using a genotypic model of association in females and an allelic model in males (who have only one copy of the X chromosome). Because genotype calling was performed blind to the gender of samples, some males appear to be heterozygous at some X chromosome SNPs. We treated all such heterozygous calls as missing. We computed missing data proportion and plate test *P*-values in males and females separately, and tested for departure from Hardy-Weinberg in female controls and for differences in frequency between males and females. We excluded SNPs with MAF < 1%, missingness > 5% (Gambia, Malawi) or 2.5% (Kenya) in males or females, or plate test $P < 0.01$ in males or females. We further excluded SNPs with Hardy-Weinberg $P < 1\times10^{-20}$ in female controls or showing significant difference in allele frequencies between males and females ($P < 1\times10^{-20}$). Extended Data Figure 2b shows the number of SNPs excluded by each criterion.

Some regions of the X chromosome showed an elevated number of male heterozygote calls, contributing to the high number of SNPs excluded due to missingness in males. These included SNPs in the pseudo-autosomal regions at either end of the chromosome[39] (indicating that XY designation in the chip manifests does not adequately cover these regions) as well as SNPs within the X transposed region near the centromere.

## Phasing and imputation

We phased genotype data in each cohort separately using SHAPEIT v2.r644[40], specifying 200 hidden states and an effective population size of 17469, as recommended for African populations by the SHAPEIT documentation, and phasing each chromosome separately. We used IMPUTE v2.3.0[41,42] to impute phased genotypes into the 1000 Genomes[6] Phase I integrated variant set (version of 24th August 2012, as downloaded from the IMPUTE website, which we refer to here as the 1000 Genomes reference panel) in 5Mb chunks with a buffer region of 500kb. For phasing and imputation we used the combined HapMap recombination map in build 37 coordinates included with the 1000 Genomes reference panel. Unless otherwise stated, downstream analyses included only SNPs with minor allele frequency > 0.5% and impute info metric > 0.75.

## Assessment of imputation

At each genotyped SNP, IMPUTE computes squared correlation (referred to here as accuracy) and concordance between typed genotypes and genotypes obtained by masking the SNP and re-imputing. To assess imputation performance, we plotted the cumulative distribution of concordance (Extended Data Figure 3a) and accuracy (Extended Data Figure 3b), accuracy by frequency (Extended Data Figure 3d) as well as average per-sample accuracy (Extended Data Figure 3c). We also assessed accuracy relative to direct typing on the Sequenom platform at variants typed in the *FREM3/GYPE*, *ARL14* and *INPP4B* regions of association as described below.

## Association testing and meta-analysis

**Association testing—**We used SNPTEST to test for association at approximately 38 million variants obtained through imputation, including 5 PCs as covariates to control for population structure separately in each cohort. SNPTEST uses a missing data likelihood to account for the uncertainty in genotypes at imputed SNPs. We fit additive, dominant, recessive, and heterozygote models of association and ran SNPTEST separately in each imputation chunk. Below, we refer to the estimated effect size for population *i* and mode of inheritance *m* as $\beta_{i,m}$ and its estimated standard error as $SE_{i,m}$.

**Frequentist meta-analysis—**For each SNP and each mode of inheritance (additive, dominant, recessive, heterozygote) we computed the fixed-effect inverse variance-weighted meta-analysis effect size, standard error, and *P*-value. In this context, fixed-effect meta-analysis assumes a single true effect size which is identical between the three cohorts, and can be thought of as finding the maximum likelihood estimate of the effect size under the assumption that the likelihood in cohort *i* is proportional to the density of the normal distribution with mean $\beta_{i,m}$ and standard error $SE_{i,m}$.

**Bayesian meta-analysis—**We have previously[1,3] used Bayesian meta-analysis techniques to allow for between-population heterogeneity of effect sizes in these three cohorts. Here, we applied this method to compute Bayes factors for association under four modes of inheritance and six different models of correlation between cohorts. In this framework, true effect sizes are modelled by a multivariate normal distribution centred on zero and with a given prior covariance matrix which can be written as $\sigma^2 P$, where $\sigma^2$ is a

prior variance controlling the magnitude of plausible effects and *P* is a prior correlation matrix.

The correlation models we used were:

*Fixed effects* (all elements of *P* equal to 1): as with frequentist fixed-effect meta-analysis, this assumes effect sizes are equal across the three cohorts.

*Correlated effects* (all off-diagonal elements of *P* equal to 0.96): this assumes effect sizes are similar but allows for some variability.

*Independent effects* (all off-diagonal elements of *P* equal to 0.1): this assumes there is little similarity between effect sizes in different cohorts.

*Structured effects* We also considered models where effects in the two East African populations (Kenya and Malawi) were more similar to each other than to that in Gambia. We assumed either effects were fixed between Kenya and Malawi and correlated ($\rho$=0.96) between East Africa and Gambia (we refer to this as the fixed-structured-effect model), or that effects were correlated ($\rho$=0.96) between Kenya and Malawi and largely independent with Gambia ($\rho$=0.1) (referred to as the correlated-structured-effect model).

We used prior variance parameters of $\sigma^2=0.2^2$, reflecting a belief in relatively small effects (odds ratio < 1.5 with 95% probability)[37], and $\sigma^2=0.75^2$, reflecting a belief in larger effects (odds ratio < 4.5 with 95% probability).

To form a single summary measure of evidence for association we formed a model-averaged Bayes factor (BF), referred to as the mean BF and denoted $BF_{avg}$, as a weighted average of model-specific Bayes factors using the following weighting scheme. We assigned weights of 0.4 for additive mode of inheritance and 0.2 for dominant, recessive, or heterozygote modes of inheritance, reflecting a belief that variants which tag causal variants by LD are more likely to display additive effects. We assigned a weight of 0.4, 0.2, and 0.1 to fixed-, correlated-, and independent-effect models, and 0.2 and 0.1 to fixed-structured and correlated-structured effects models. Finally we assigned weight of 0.5 to small-effect models ($\sigma^2=0.2^2$) and large-effect models ($\sigma^2=0.75^2$). Overall prior weights were assigned by multiplying across these categories; for example, the model representing small effect size distribution, fixed-effect across cohorts with additive mode of inheritance (denoted small-fix-add) was assigned prior weight equal to $0.5\times0.4\times0.4=0.08$, while the model representing small, correlated-structured, dominant effects (denoted small-cor-str-dom) was assigned a prior weight of $0.5\times0.1\times0.2=0.01$. For each SNP we also recorded the model having the highest posterior weight, and refer to its Bayes factor as the max BF (denoted $BF_{max}$). Extended Data Figure 5a depicts slices through the combined prior on effect sizes across three cohorts for additive effect models. The mean BF behaves similarly to a minimum over all four fixed-effect meta-analysis *P*-values, but additionally captures effects that vary between cohorts (Extended Data Figure 5b).

As described previously[1,3], to computed model-specific Bayes factors efficiently we used an approximation of the likelihood by the density of a normal distribution with the estimated

mean ($\beta_{i,m}$) and standard error ($SE_{i,m}$) in each cohort. Thus, overall, observed effect sizes are modelled as normally distributed around zero with covariance that depends on the prior covariance in true effect sizes and on model standard errors,

$$(\beta_{i,m}) \sim N(0, \sigma^2 P + V)$$

where $V$ is a diagonal matrix with $i$th diagonal entry equal to the squared standard error, $SE^2_{i,m}$. The approximate or asymptotic Bayes factor can then be computed by evaluating a ratio of two normal densities. In the univariate case this method is the same as described previously by Wakefield[43]. To facilitate working with genome-wide meta-analysis results, we wrote custom software to compute frequentist and Bayesian meta-analyses, and stored details of genotype counts, association model fit, and meta-analysis results directly in a SQLite database file.

Further discussion of the Bayesian approach can be found in Supplementary Note 4.

## X chromosome association testing and meta-analysis

Association testing on the X chromosome was performed as for the autosomes with a few differences as follows. We ran SNPTEST separately in males and females, estimating effect sizes under additive, dominant, recessive and heterozygote modes of inheritance in females. In this usage, SNPTEST assumes a model of complete inactivation so encodes male genotypes as 0/1 and females as 0/0.5/1 for an additive mode of inheritance. We than meta-analysed the six gender-specific association analyses to produce frequentist fixed-effect and model-averaged bayesian meta-analyses.

For Bayesian analysis, in addition to summing over models of between-population heterogeneity, we adopted the view that differences in sex might lead to heterogeneity in effect. We therefore included models of heterogeneity between males and females as follows. Let $\rho_{sex}$ denote the correlation between effects in male and female samples within a single population. We included models where males and females have the same effect ($\rho_{sex}$=1, termed fixed-sex model and given prior weight 0.45), correlated effects ($\rho_{sex}$=0.96, termed correlated-sex model and given prior weight 0.225) or essentially independent effects ($\rho_{sex}$=0.1, termed independent-sex model and given prior weight 0.225). Because some parts of the X chromosome escape inactivation[39,44], we also included a model where the effect in females is twice that in males (given prior weight 0.1).

To fully specify prior correlation for each model, for each pair of populations *A,B* we also need to specify the correlation (denoted $\rho_{cross}$) between males in A and females in B. We set $\rho_{cross} = \rho_{pop} \times \rho_{sex}$ where $\rho_{pop}$ is the chosen prior correlation between same-sex samples in *A* and *B* (i.e. $\rho_{pop}$ =1, 0.96 or 0.1 as defined above).

As above, we formed overall weights by multiplying across categories, so that for example the model of small, additive effects that are fixed across populations and across sexes (denoted small-fixsex-fix-add) had prior weight 0.5×0.4×0.4×0.45=0.036.

### Linear mixed model analysis

To compare association test results for logistic regression as implemented in SNPTEST with the use of a linear mixed model, we reran association test scans in each discovery cohort using the program MMM[45]. We used the same genome-wide relatedness matrix as used to compute principal components (see above) and assumed an additive model of association. We plotted the –log10( *P*-value ) for MMM against the corresponding –log10( *P*-value ) based on the SNPTEST scan (which used 5 PCs as described above), for each discovery population and for fixed-effect meta-analysis (Extended Data Figure 4f). *P*-values under both methods at tier 1 loci are listed in Supplementary Table 1.

### Lead SNPs, region and tier definitions

We formed a list of lead SNPs within approximately independent regions of interest as follows. We restricted to variants with IMPUTE info measure at least 0.75 across all three populations and ranked variants by the model-averaged Bayes factor (highest to lowest). We iteratively picked lead SNPs from the top of the list and excluded other variants within a recombination interval of 0.25cM±25kb centred at the lead SNP (referred to below as the association region), continuing until no more SNPs remained with $BF_{avg}$>250 or $BF_{max}$>2500.

We grouped lead SNPs into two tiers as follows:

- tier 1: all lead SNPs with $BF_{avg}$>2500 or $BF_{max}$>25000
- tier 2: all lead SNPs not in tier 1 with $BF_{avg}$>1000 or $BF_{max}$>10000

In total, across the autosomes and the X chromosome there were 34 regions in tier 1 and 73 in tier 2, with association regions covering approximately 13Mb and 26Mb of the genome respectively.

### Regional association analysis

For each tier 1 and 2 region, we examined the pattern of association in the region, generating a regional association plot for the region annotated with details of the meta-analysis for the lead SNP as follows.

**LD computation—**In each region we computed pairwise LD statistics between the lead SNP and surrounding SNPs using best-guess imputed haplotypes for control samples across three populations. To facilitate this computation, we stored imputed haplotypes in a SQLite-format database allowing us effective random access to haplotype data. We used R to compute both Pearson correlation coefficient ($r^2$) and Lewontin's |D'| between the lead SNP and all other SNPs in the region.

**Regional association plot—**For each SNP in tier 1 and 2 we plotted log10($BF_{avg}$) in the association region around the lead SNP plus 1Mb on either side, colouring points according to LD, with outer circles representing $r^2$ and inner circles representing |D'|. We further annotated SNPs that were typed in at least one cohort (using black plusses) and SNPs that had Sequenom genotype data available (with black triangles).

**Regional annotations—**We plotted all ABP variants (Supplementary Tables 4 and 5 of Leffler *et al*[4]), eQTLs from the Genotype-Tissue Expression project[46] (GTEx), and previously reported[47] GWAS loci. Where SNPs matched an imputed or typed variant, we computed LD statistics and coloured these points as above. We plotted all RefSeq genes (as downloaded from the UCSC Genome Browser MySQL database[48] on 2013-03-18) in the region, annotating the direction of transcription. We further plotted local recombination rate estimates from the HapMap combined recombination map included with the IMPUTE haplotypes described above.

**Meta-analysis—**We annotated plots with effect sizes and confidence intervals for the lead SNP for each mode of inheritance (additive, dominant, recessive and heterozygote) that informed the model averaging. We also produced barplots showing the posterior distribution on models of association using the prior weights described above.

**Cluster plots—**Finally, for each region, we produced and inspected cluster plots for all typed SNPs with $BF_{avg}$ >= 10 that were not excluded by QC. As phasing fills in missing genotypes and potentially improves genotype calling based on LD with surrounding SNPs, we coloured plots based on genotype calls taken from the output of phasing.

**Website—**Regional association plots and cluster plots can be viewed online at http://www.malariagen.net/resource/14.

## Validation and replication typing

Based on a preliminary version of the data presented here we selected SNPs for typing on the Sequenom platform across the whole MalariaGEN Consortial Project 1 sample set, which includes the discovery samples, further cases and controls in the same populations that were not included in the GWAS, and further large sample sets from five other populations from sub-Saharan Africa (see Extended Data Figure 1a). Data were available for SNPs tagging the lead markers in the *FREM3/GYPE, INPP4B* and *ARL14* regions ($r^2$>0.5 in controls, as estimated using the EM algorithm) as well as other regions not represented in tier 1.

## Replication analysis

Replication analysis using Sequenom data was restricted to the set of samples with less than 10% missingness as measured across the set of 70 SNPs chosen for replication typing. In each population we conducted logistic regression in R, including five principal components (discovery samples) or reported ethnicity (replication samples) as covariates to control for population structure. For each GWAS lead SNP, we examined each Sequenom SNP in the region and computed $r^2$ and |D'| with the lead SNP. To allow for the effects of incomplete LD on the Bayesian model fitting, we recomputed the Bayesian analysis in the discovery samples based on Sequenom genotypes at each SNP, to obtain a Sequenom-based mean Bayes factor and a 'best model' Bayes factor at the model with highest posterior weight. Where LD is incomplete or where imputation is imperfect, this model may differ from the best model for imputed data.

For each SNP we computed fixed-effect meta-analysis across discovery samples, across replication samples, and across all samples. We also computed the Bayes factor for replication samples ($BF_{replication}$) for the model with highest posterior weight in the discovery samples. To compute an overall Bayes factor for association, we combined the discovery $BF_{avg}$ computed at the imputed lead marker with the replication bayes factor at the Sequenom SNP, as $BF_{overall} = BF_{avg}$ x $BF_{replication}$. We use the imputed lead marker here because the number of discovery samples directly typed was smaller than the number of imputed samples. Conditional on the lead imputed and replication markers reflecting the same signal of association, this $BF_{overall}$ represents an overall measure of the evidence for association at the locus that reflects all the samples in our study.

A discussion of the replication evidence in the *FREM3/GYPE, INPP4B* and *ARL14* regions can be found in Supplementary Note 1.

MalariaGEN encourages individual study sites to perform more detailed analyses of local patterns of disease association, and a Tanzanian-focused analysis of *FREM3* and other candidate SNPs that were genotyped as part of this study is reported elsewhere[49].

## Credible interval analysis

In a given region of the genome, under the assumption that exactly one variant (that is accessible to our typing or imputation) is causal, the posterior probability that each variant is the causal variant can be computed by a simple reweighting of Bayes factors[50]. Using this approach we computed 95% and 99% credible intervals (i.e. the smallest set of SNPs accounting for 95% or 99% of the posterior mass) for variants in the association region around rs184895969, plus a margin of 50kb at either end (Extended Data Figure 7b). We noted that rs181620317, which is annotated as a missense mutation for the gene FREM3, is within the 95% confidence interval in our data. We note that while this analysis is simple and appealing, its interpretation depends on assumptions about the true disease model, and on the behaviour of imputation in the region[50]; in particular the difficulty of imputing variants around the three glycophorin genes (Extended Data Figure 8) might make this analysis fail to capture putatively important variation within or around *GYPA, GYPB*, or *GYPE*.

## Sequence homology and alignability in the glycophorin region

To investigate the location of our GWAS signal with respect to the pattern of sequence homology around the three glycophorin genes, we generated a dot plot (Extended Data Figure 8a) showing co-occurrence of k-mers in the human reference sequence[51] in the region. Considerable sequence homology was observed over a region of about 350kb covering the three genes. Our lead GWAS marker and the ABPs lie just outside this region.

The high level of homology should affect our ability to align probes or sequences to this part of the genome. To confirm this, we also plotted the UCSC alignability track (‘CRG alignability 100’, Extended Data Figure 8c), which shows the degree to which the 100-mer starting at each position in the reference sequence is alignable (with a value of 1/n indicating that the 100-mer aligns to n positions across the genome, allowing up to two mismatches).

As expected, even short regions of shared kmers affect alignability considerably, with alignability dropping to an average of about 0.7 within the large region of homology. Similarly, we plotted imputation performance (as measured by the IMPUTE info score) and observed a marked drop within the region of sequence homology.

## Structural variation in the glycophorin region

We attempted to identify structural variation in the glycophorin region by examining sequence read data generated by the 1000 Genomes' project, using the set of BAM files available from the 1000G project in October 2014, downloaded from ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data. These data contain a mixture of read lengths, with most reads of at least 90bp and some reads of 76bp. We considered only reads with mappinq quality at least 20. For each sample we computed coverage across the glycophorin region, normalised by average coverage computed across chromosome 1, and refer to these values here as 'coverage'.

Coverage values across the region were correlated with genome alignability (as defined above), with coverage dropping substantially in regions of low alignability. We therefore restricted attention to the set of perfectly alignable positions, here defined as positions such that every 100-mer overlapping the position in the reference sequence aligns uniquely, allowing up to two mismatches (and computed using the CRG alignability 100 track described above). As expected, coverage at perfectly alignable positions showed little or none of the variation present in genome alignability (Extended Data Figure 8e–g).

Two large structural variants, present at frequencies of at least 1% in LWK+YRI, were evident in coverage data. Genotypes for both these variants were called by the 1000 Genomes phase I and are referred to as esv2668125 (also called MERGED_DEL_2_26708) and esv2662558 (also called MERGED_DEL_2_26722). Both deletions putatively represent deletions of all or part of *GYPB*. We noted some uncertainty as to the location of the breakpoints of both deletions - with 1000 Genomes' breakpoints differing from the breakpoints as they appeared in coverage data by at least 10kb. We further noted that three samples (NA18519, NA19185, NA19222) that were called as heterozygote for one or both deletions by the 1000 Genomes project, appeared to have homozygous genotypes (red, blue and green lines, Extended Data Figure 8e–g).

## Expression QTL analysis

To investigate the effect of associated SNPs in the *FREM3* region on expression of nearby genes, we downloaded publically available data on RNA expression levels[17] in HapMap samples, and plotted expression levels of genes in the region (Figure 1) against genotypes at the lead SNP and other SNPs of interest in African samples (YRI+LWK, Extended Data Figure 9). Data was available for genes *INPP4B, USP38, GAB1, SMARCA5, GYPE, GYPB, GYPA, HHIP, ANAPC10 and ABCE1*, with three probes available for *GYPE*, two for *GAB1* and one probe for each of the other genes. Five samples in this dataset (NA19318, NA19324, NA19377, NA19429, NA19190) carry the protective allele at the lead marker rs184895969, while only two carry the protective allele at the directly-typed rs186873296. For each gene and SNP we tested for a trend of genotype on expression using linear

regression. Extended Data 9 shows all probes for the glycophorins as well as other regional genes for which a *P*-value less than 0.05 was observed.

### GENECLUSTER analysis

We used the program TREESIM[21] to estimate an ancestral recombination graph for the LWK and YRI individuals in the 1000 Genomes Project haplotype data in a region from 144.6Mb to 144.8Mb on chromosome 4 centred on the lead SNP in the *FREM3 / GYPE* region. We then ran GENECLUSTER[21] in the region, allowing it to assign either one or two causal mutations in each marginal genealogy. GENECLUSTER works by probabilistically assigning study individuals to the tips of the tree estimated by TREESIM, and attempts to explain case/control status by assigning causal mutations to the branches of the tree in a Bayesian framework. This analysis is somewhat similar in spirit to our marginal SNP analysis, but has important differences. In particular, GENECLUSTER may detect associated mutations anywhere on the tree (which may not correspond directly to any typed or imputed variant), and can assess models of association involving more than a single causal variant. However, GENECLUSTER does not take into account principal components or otherwise control for population structure, and for computational reasons we only included the Luhyan (LWK) and Yoruban (YRI) populations from the 1000 Genomes reference panel in the analysis.

To investigate the relationship between ancestry and variants of interest, we plotted IMPUTE reference panel haplotypes in the region ordered by the marginal tree at the position of the imputed lead marker (rs184895969), annotating the lead imputed and Sequenom-typed markers, ABPs, previously reported *GYPE* eQTLs[52], common deletions, and variants determining the M/N and S/s blood groups as described in Supplementary Note 2 (Figure 3).

### Analysis of enrichment of malaria-associated loci in functional categories

Full details of enrichment analyses are provided in Supplementary Note 3.

### Code availability

The SNPTEST software for genome-wide association testing is available at https://mathgen.stats.ox.ac.uk/genetics_software/snptest/snptest.html.

Executables and source code for inthinnerator are available at http://www.well.ox.ac.uk/~gav/inthinnerator.

Further software for generating key results for this paper will be made available at http://www.malariagen.net/resource/14.

## Supplementary Material

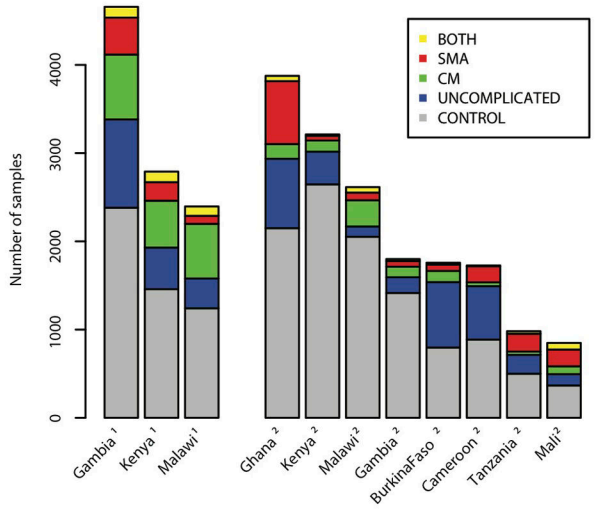
Refer to Web version on PubMed Central for supplementary material.

# Extended Data

**a**

| Cohort | Site | Ethics approving institution | Ethnics committee | Genotyping platforms |
|---|---|---|---|---|
| | | ***Discovery and replication*** | | |
| Gambia | Medical Research Council Unit, The Gambia | MRC Gambia and Gambia Government | MRC/Gambia Government Ethics Committee | Illumina HumanOmni2.5-8, Sequenom |
| Malawi | Malawi-Liverpool Wellcome Clinical Research Programme, Blantyre | University of Malawi, College of Medicine | CoM REC | Illumina HumanOmni2.5-8, Sequenom |
| Kenya | KEMRI-Wellcome Research Programme, Kilifi | KEMRI, Kilifi | KEMRI REC | Illumina HumanOmni2.5-4, Sequenom |
| | | ***Replication only*** | | |
| Mali | University of Bamako, Bamako, Mali. | University of Bamako, FMPOS, MRTC | FMPOS Institutional Ethics Committee | Sequenom |
| Burkina Faso | Centre National de Recherche et de Formation sur le Paludisme | Ministry of Health & Ministry of Science and Education | Health Research Ethics Committee | Sequenom |
| Ghana | Navrongo Memorial Institute for Medical Research with Navrongo Health Research Centre | Navrongo | Navrongo IRB | |
| | Kwame Nkrumah University of Science and Technology, Kumasi | Ghana Health Service | GHS ERC | Sequenom |
| | | School Medical Sciences, KNUST | Committee on Human Research Publication and Ethics | |
| Cameroon | University of Buea | University of Buea | IRB | Sequenom |
| | | Govt. of Cameroon | Provincial Delegate for Public Health | |
| Tanzania | Joint Malaria Programme, Kilimanjaro Christian Medical Centre | London School of Hygiene and Tropical Medicine | LSHTM ERC | Sequenom |
| | | | NIMR Research Coordinating Committee | |
| | | National Institute for Medical Research (NIMR), Tanzania | | |

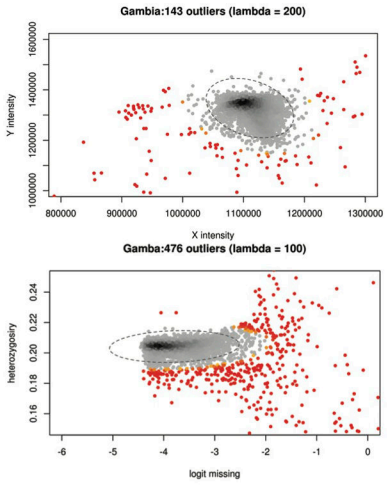
**b**



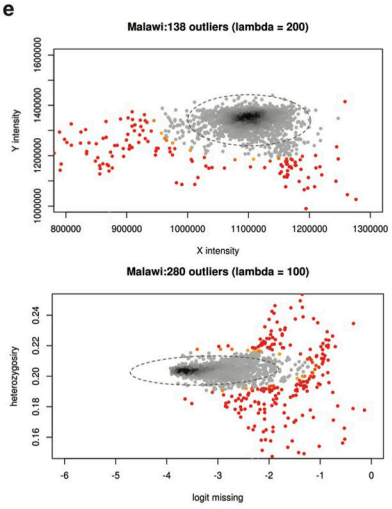

**c**

| | Samples genotyped | Missingness / heterozygosity / intensity outlier | Removed as related | Total after sample QC (cases/controls) |
|---|---|---|---|---|
| Gambia | 5596 | 490 (396*) | 279 | 4921 (2430/2491) |
| Malawi | 3088 | 327 | 245 | 2516 (1194/1322) |
| Kenya | 3869 | 254 | 473 | 3142 (1506/1478)† |

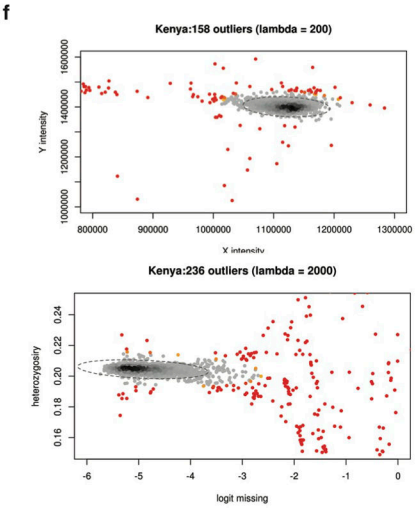
**d**




**e**




**f**




**Extended Data Figure 1.**

Sample collections included in the study. **a**) Study sites and ethics approving institutions. **b**) Phenotypic makeup of discovery and replication samples from each site. 'UNCOMPLICATED' refers to case individuals who were not identified as cerebral malaria (CM) or severe malarial anaemia (SMA) cases. 'BOTH" refers to individuals who have both CM and SMA phenotypes. **c**) Overall sample counts and number of samples excluded by

each QC criterion. (*) denotes the number of samples removed after explicitly including samples with low heterozygosity in Gambia. (†) The Kenyan cohort included parents of a subset of case samples; these were not used in subsequent analyses. **d**) Plots of average genome-wide heterozygosity and missingness with outliers coloured, as output by the ABERRANT algorithm.

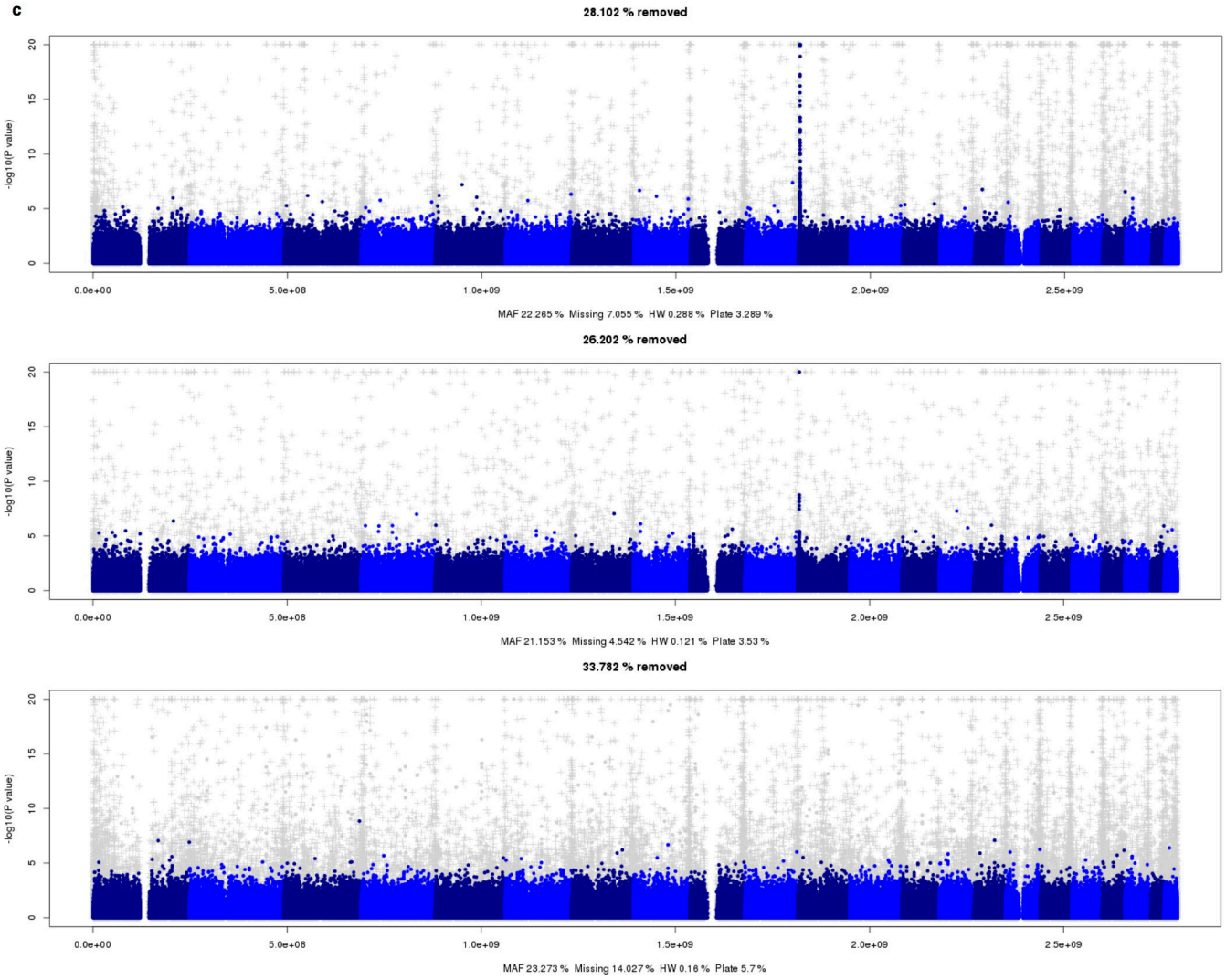
**a**

| | Total pre-QC | MAF | Missingness | HWE | Plate | Cluster plot | Total remaining |
|---|---|---|---|---|---|---|---|
| Gambia | 2269360 | 505275 | 160093 | 6528 | 74628 | 58 | 1631558 (72%) |
| Kenya | 2269360 | 480031 | 103064 | 2751 | 80114 | 67 | 1674680 (74%) |
| Malawi | 2269360 | 528137 | 318314 | 3622 | 129358 | 114 | 1502608 (66%) |

**b**

| | Total pre-QC | MAF | Missingness (females) | HWE | Plate females | Missingness (males) | Plate males | Diff | Total remaining |
|---|---|---|---|---|---|---|---|---|---|
| Gambia | 52181 | 10011 | 2180 | 154 | 1183 | 6403 | 1483 | 164 | 35474 (68%) |
| Kenya | 52181 | 9909 | 2083 | 69 | 827 | 7499 | 1249 | 0 | 34312 (66%) |
| Malawi | 52181 | 10830 | 5281 | 206 | 1365 | 11192 | 1761 | 41 | 31550 (60%) |

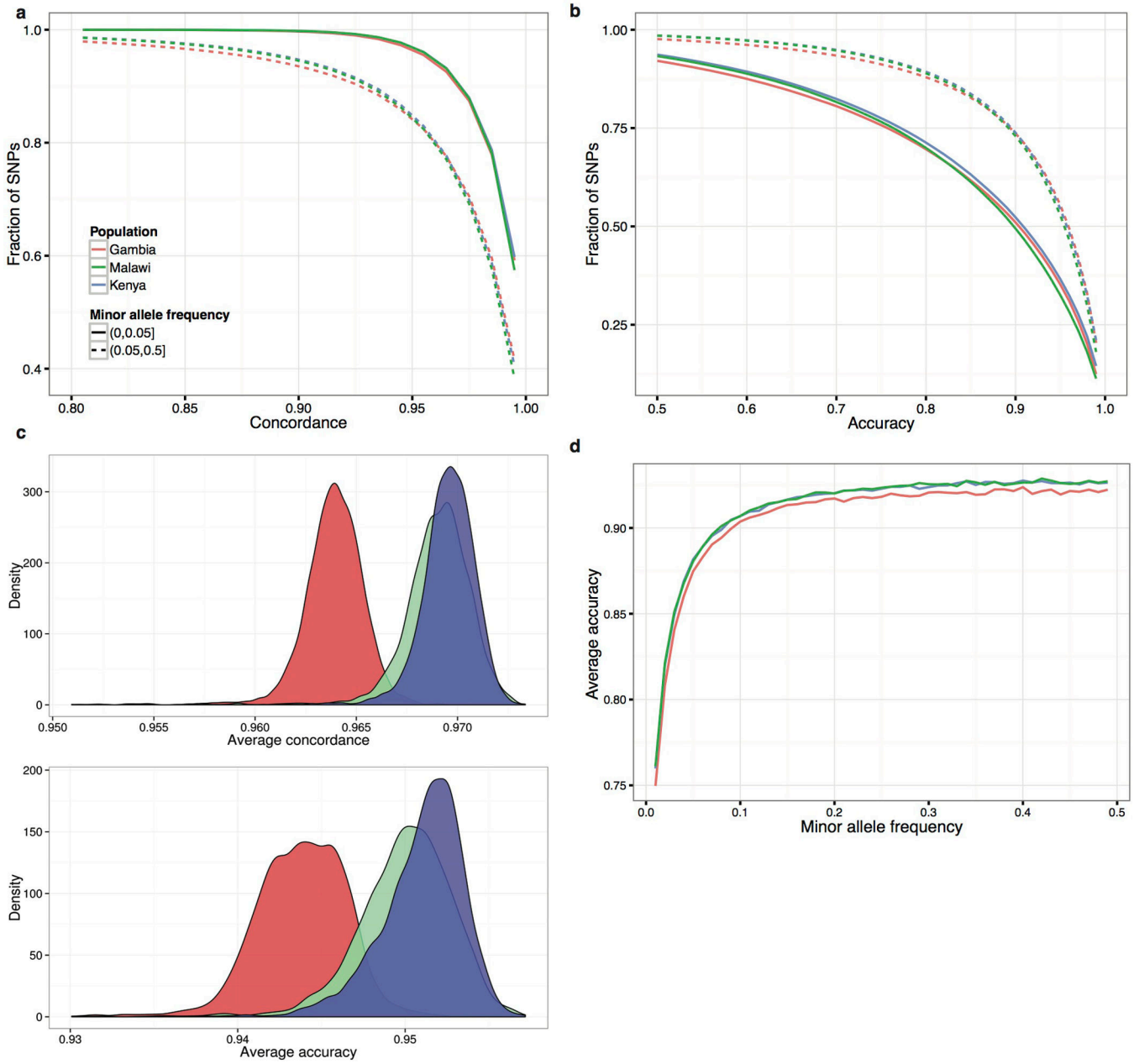
**c**




**Extended Data Figure 2.**

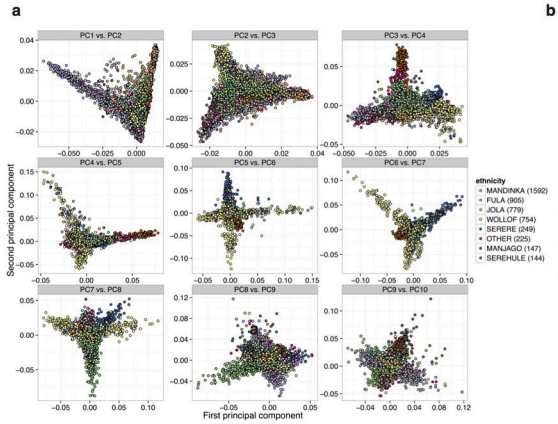
Genotyped SNP quality control (QC) for the 3 discovery cohorts. **a,b**) Total numbers of pre- and post-QC SNPs on **a**) the autosomes and **b**) the X chromosome, and numbers of SNPs

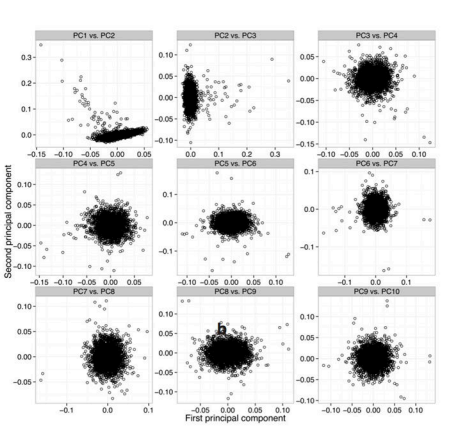
excluded by each QC criteria. MAF refers to minor allele frequency, HWE to Hardy-Weinberg equilibrium, Plate to the plate test of association and Diff. to the test of difference in frequency between males and females. Details of QC are given in Methods. **c**) Plot showing the –log10(*P* values) for the genotypic association test in the discovery data including the first 5 principal components as covariates. Grey dots show SNPs that are removed due to the QC as defined in Methods. The total fraction of SNPs removed from each cohort is given at the top of the plot.

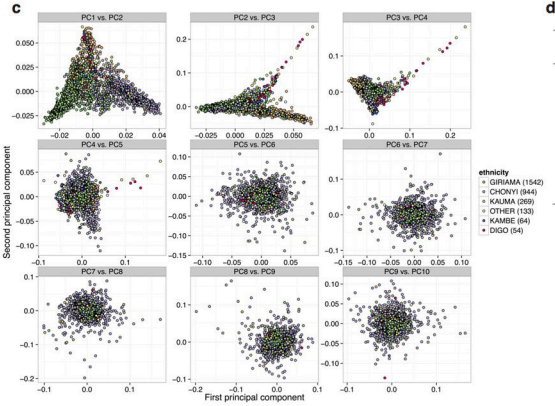



**Extended Data Figure 3.**

Imputation performance. **a,b**) Empirical distribution of concordance and accuracy ($r^2$) between typed and re-imputed SNPs in the three discovery cohorts. Solid lines represent SNPs with frequency below 5% and dashed lines represent SNPs with frequency of at least 5%. **c**) Per-sample concordance and accuracy (type 0 $r^2$) across the whole genome, as estimated by reimputing genotyped SNPs. Values are averaged over imputation chunks. **d**) Average accuracy between genotype and re-imputed SNPs in each cohort, plotted against frequency, in 1% frequency bins.

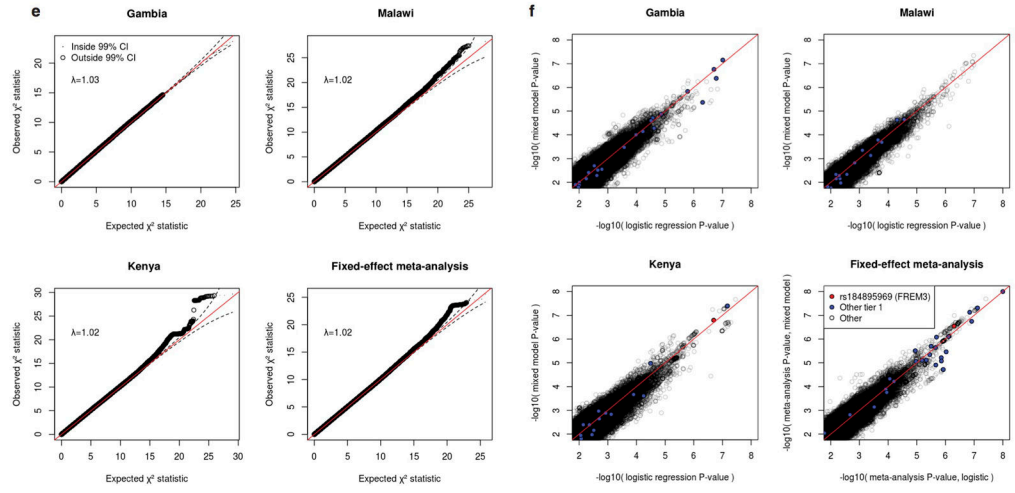


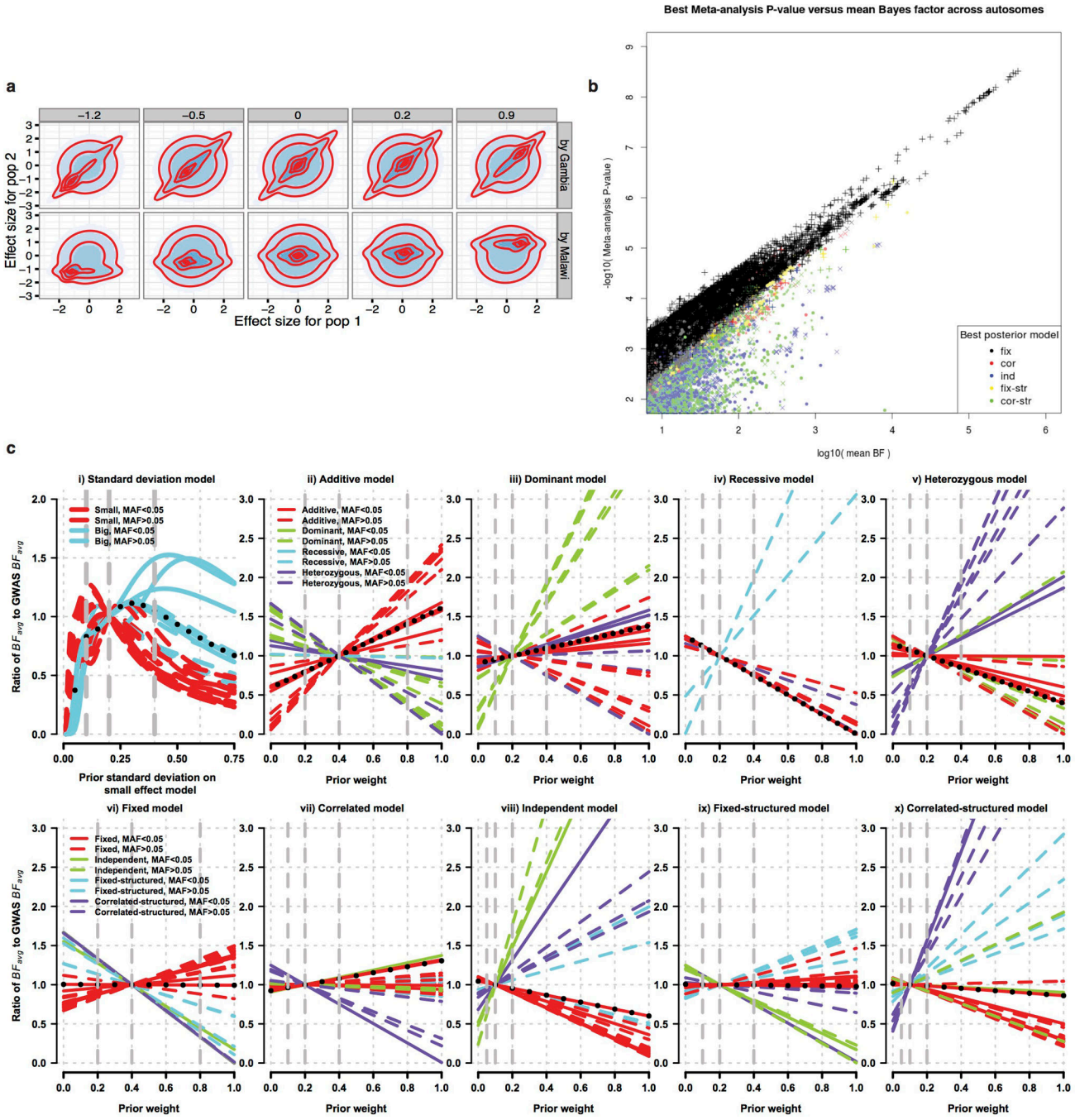

d

| | Gambia | | Malawi | | Kenya | |
|---|---|---|---|---|---|---|
| PC | P-value | Effect direction | P-value | Effect direction | P-value | Effect direction |
| 1 | 5.43E-21 | + | 6.28E-03 | + | 1.65E-21 | - |
| 2 | 2.11E-16 | + | 5.27E-02 | + | 3.02E-01 | - |
| 3 | 1.76E-02 | + | 1.31E-01 | + | 6.56E-04 | + |
| 4 | 4.45E-11 | + | 7.52E-01 | + | 8.84E-04 | - |
| 5 | 3.93E-01 | - | 7.76E-01 | + | 1.84E-02 | + |
| 6 | 8.00E-01 | + | 6.14E-01 | - | 2.71E-02 | - |
| 7 | 3.20E-01 | + | 1.56E-01 | - | 4.10E-01 | - |
| 8 | 6.20E-02 | + | 4.32E-01 | + | 1.34E-01 | + |
| 9 | 2.17E-02 | + | 4.92E-01 | - | 8.36E-01 | + |
| 10 | 4.89E-11 | - | 4.31E-01 | + | 4.40E-01 | + |



**Extended Data Figure 4.**

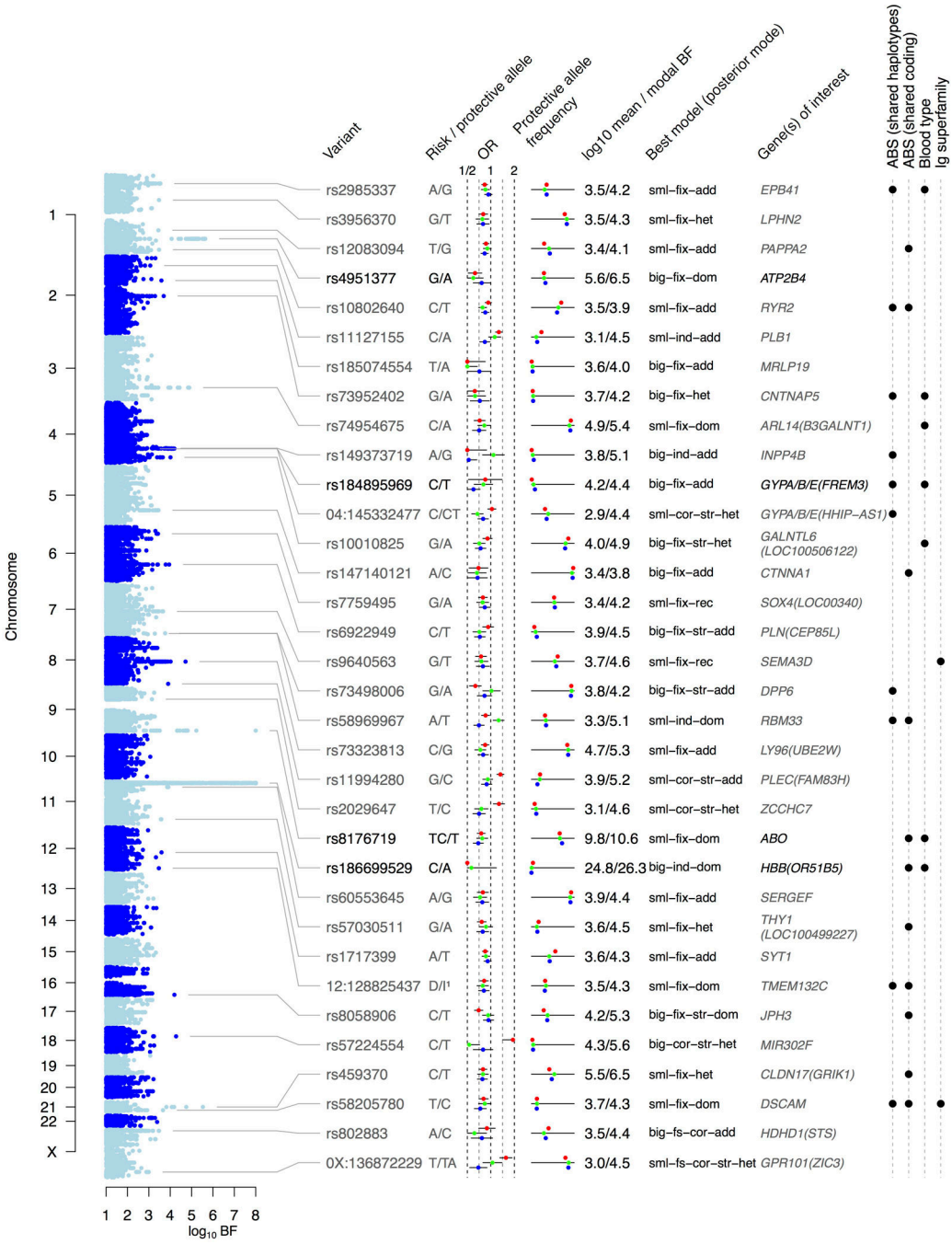
Top ten principal components (PCs) in **a**) Gambia, **b**) Malawi and **c**) Kenya. Where ethnicity was reported, points are coloured by ethnicity for ethnicities with at least 50 samples. **d**) Logistic regression *P*-values and direction of effect for the top ten principal components on Severe Malaria status in each cohort. **e**) qq-plots for additive model association test *P*-values in Gambia, Malawi, Kenya, and for fixed-effect meta-analysis. Dashed lines represent the 99% confidence interval computed marginally at each variant. Circles and points represent points lying respectively outside and inside the 99% confidence interval. **f**) Comparison of association test P-values for logistic regression (SNPTEST, x-axis) and linear mixed model (MMM, y-axis) for Gambia, Malawi, Kenya, and for fixed-effect meta-analysis. Variants in tier 1 are coloured blue, with the lead marker at the *FREM3/GYPE* region coloured red.




**Extended Data Figure 5.**

Detail of Bayesian analysis of discovery cohorts. **a**) Visualisation of slices through the combined prior on effect sizes in three cohorts for mode-of-inheritance-specific models. Top row: slices through the prior effect size on Kenya (x-axis) and Malawi (y-axis) for constant effect size in Gambia (panels). Bottom row: slices through the prior effect size on Kenya (x-axis) and Gambia (y-axis) for constant effect size in Malawi (panels). Red lines represent a factor of 10 in the prior density. **b**) Comparison of $BF_{avg}$ (x-axis) with the minimum fixed-effect meta-analysis P-value minimized across additive, dominant, recessive or heterozygote

modes of inheritance (y-axis). Values are plotted on log10 and –log10 scales. Colour indicates the heterogeneity model of the model with the highest posterior weight. **c**) Sensitivity of $BF_{avg}$ to changes in prior. Plots show $BF_{avg}$ ratio (y-axis) plotted against one-dimensional parameterisations of the prior (x-axis), for the 32 autosomal SNPs in tier 1. Solid lines represent variants with minor allele frequency < 5% averaged across populations, and dashed lines variants with minor allele frequency >= 5%. Black dots indicate the lead marker at the *FREM3/GYPE* locus. Colour indicates the effect size, mode of inheritance, or heterogeneity model for the model with highest posterior weight under the GWAS prior. Dashed grey vertical lines indicate the x-axis value corresponding to the prior used in the GWAS, and one-half and twice that value. Plots are parameterised by i) the prior standard deviation of the small-effect model keeping the prior standard deviation of the large and small-effect models in the ratio 0.75:0.2; ii-v) the prior weight on additive, dominant, recessive or heterozygote modes of inheritance; vi-x) the prior weight on fixed, correlated, independent, fixed-structured and correlated-structured models. For each parameterisation prior weights on other models are kept in the same relative proportion. For further details see Supplementary Note 4.




**Extended Data Figure 6.**

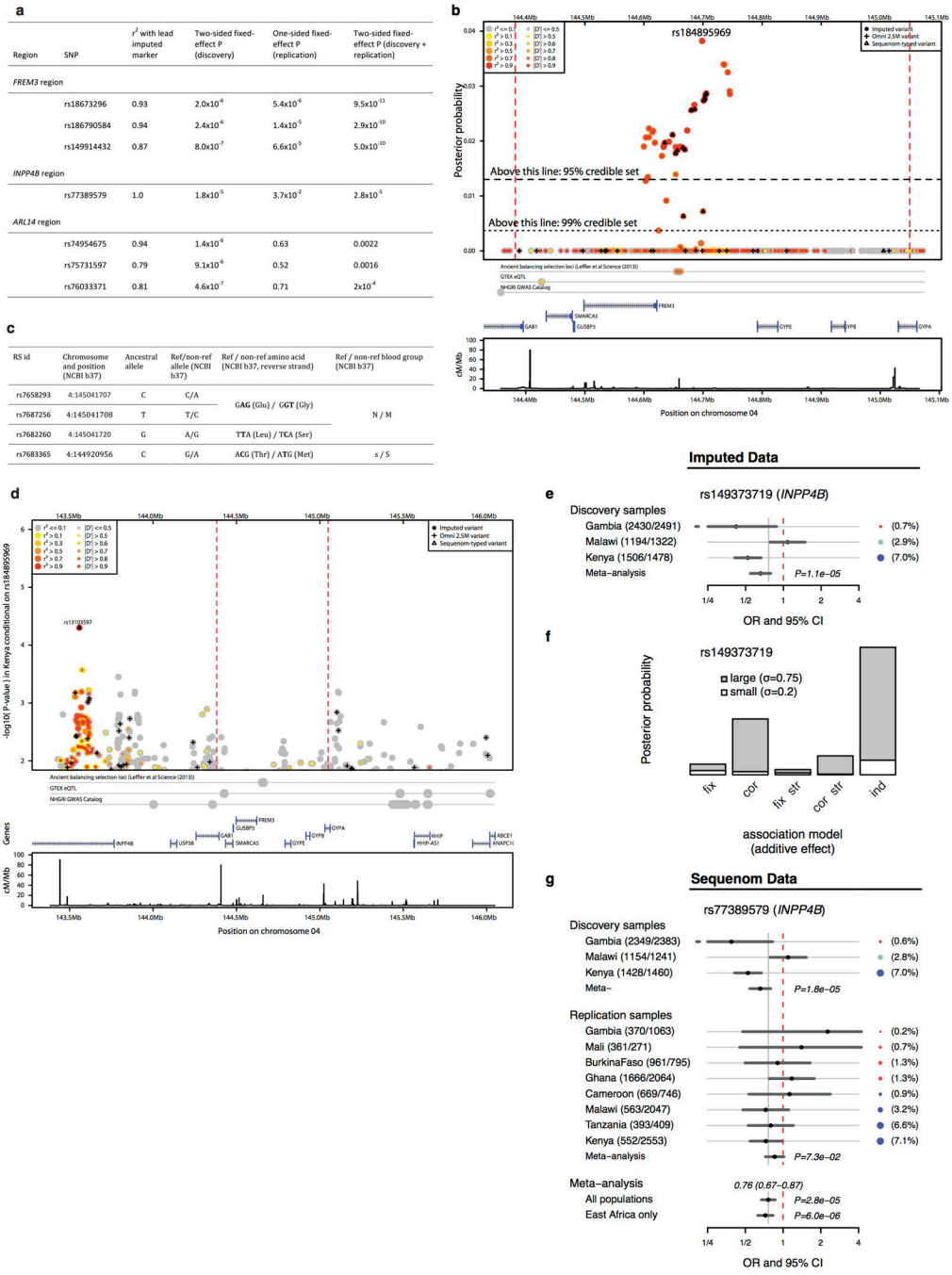
Strongest regions of association in the Bayesian analysis of the three discovery cohorts. Plot on left shows the log10 model- averaged Bayes Factor ($BF_{avg}$). Table shows the SNP with the highest $BF_{avg}$ in each region (lead SNP), gene(s) of interest in the region, the model with the highest posterior weight at the lead SNP and its BF. Coloured points indicate the odds ratio (OR) and the protective allele frequency in Gambia (red), Malawi (green) and Kenya (Blue). The right hand columns indicate regions containing shared chimp-human haplotypes or coding SNPs[4] (ABPs), blood group genes, or Immunoglobulin superfamily genes.

**a**

| Region | SNP | $r^2$ with lead imputed marker | Two-sided fixed-effect P (discovery) | One-sided fixed-effect P (replication) | Two-sided fixed-effect P (discovery + replication) |
|---|---|---|---|---|---|
| *FREM3* region | | | | | |
| | rs18673296 | 0.93 | $2.0x10^{-6}$ | $5.4x10^{-6}$ | $9.5x10^{-11}$ |
| | rs186790584 | 0.94 | $2.4x10^{-6}$ | $1.4x10^{-5}$ | $2.9x10^{-10}$ |
| | rs149914432 | 0.87 | $8.0x10^{-7}$ | $6.6x10^{-5}$ | $5.0x10^{-10}$ |
| *INPP4B* region | | | | | |
| | rs77389579 | 1.0 | $1.8x10^{-5}$ | $3.7x10^{-3}$ | $2.8x10^{-5}$ |
| *ARL14* region | | | | | |
| | rs74954675 | 0.94 | $1.4x10^{-6}$ | 0.63 | 0.0022 |
| | rs75731597 | 0.79 | $9.1x10^{-6}$ | 0.52 | 0.0016 |
| | rs76033371 | 0.81 | $4.6x10^{-7}$ | 0.71 | $2x10^{-4}$ |

**c**

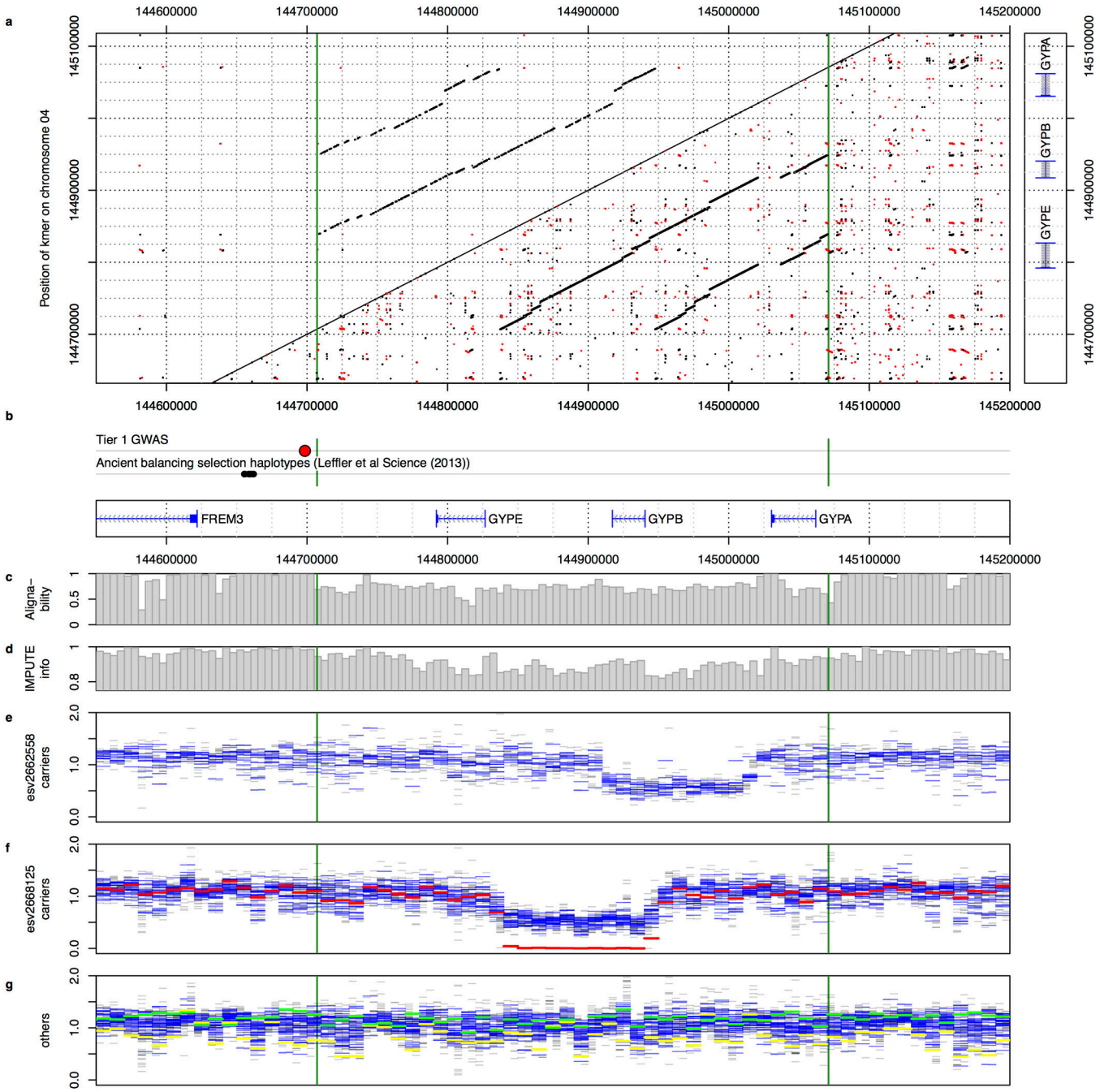
| RS id | Chromosome and position (NCBI b37) | Ancestral allele | Ref/non-ref allele (NCBI b37) | Ref / non-ref amino acid (NCBI b37, reverse strand) | Ref / non-ref blood group (NCBI b37) |
|---|---|---|---|---|---|
| rs7658293 | 4:145041707 | C | C/A | GAG (Glu) / GGT (Gly) | N / M |
| rs7687256 | 4:145041708 | T | T/C | | |
| rs7682260 | 4:145041720 | G | A/G | TTA (Leu) / TCA (Ser) | |
| rs7683365 | 4:144920956 | C | G/A | ACG (Thr) / ATG (Met) | s / S |




**Extended Data Figure 7.**

**a**) Evidence for association at directly-typed SNPs in the *FREM3/GYPE, INPP4B* and *ARL14* regions. **b**) Posterior probability that variants in the *FREM3/GYPE* region are causal assuming a single variant in the region is causal[50], based on the $BF_{avg}$ for typed and imputed variants. Dashed lines indicate the 95% and 99% credible sets. See Figure 1 legend for further details. **c**) Details of SNPs encoding the common MNS blood groups. Coordinates and alleles are with respect to the NCBI b37 human reference sequence. **d**) Evidence for possible independence of effects at the *FREM3* and *INPP4B* loci in Kenya by conditional

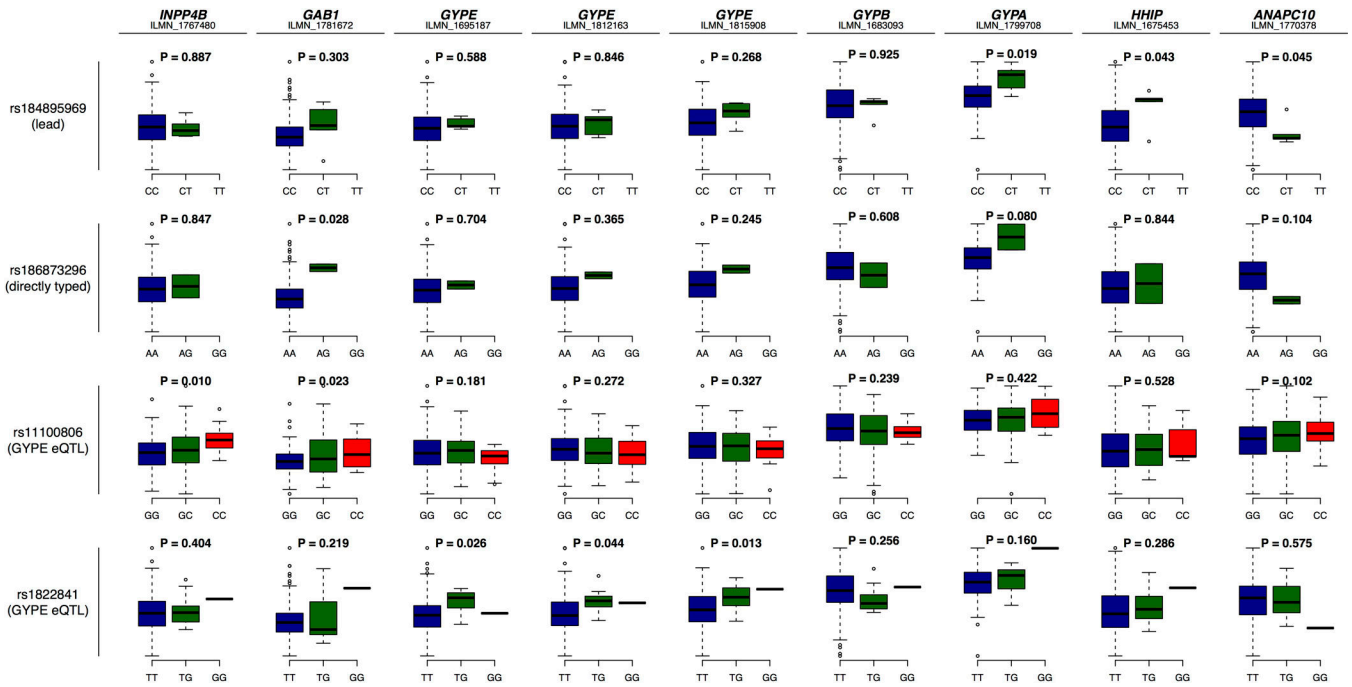
analysis. Y-axis represents -log10(association *P*-value) conditional on the imputed dosage at rs184895969. Points are coloured by LD with the top SNP rs13103597. **e**) Forest plot showing sample size, estimated odds ratio and 95% confidence interval for the lead imputed SNP (rs149373719) in *INPP4B* under an additive model of association. **f**) Bar plot showing the posterior weight on different models of heterogeneity at rs149373719 under the prior used in the GWAS, assuming an additive model of association. **g**) Forest plot showing evidence in both discovery and replication samples in the Sequenom data at rs77389579 in *INPP4B*. See Figure 2 legend for further details.




**Extended Data Figure 8.**

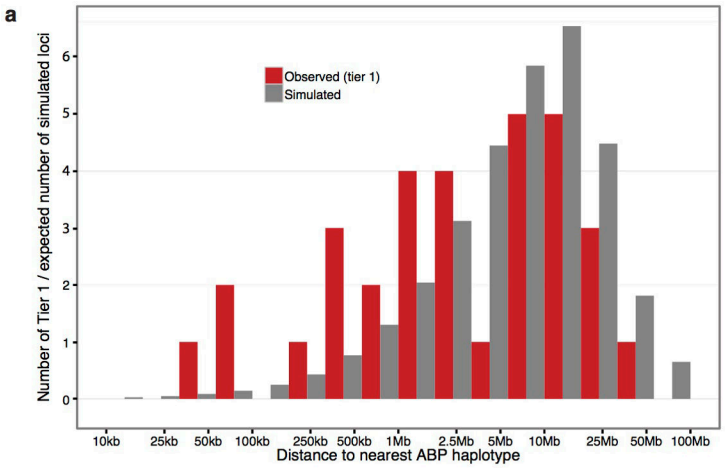
Sequence homology, alignability and structural variation in the glycophorin region. **a**) co-occurrence of 100-mers (upper triangle) and 25-mers (lower triangle) in the human reference sequence. Each point represents a kmer that maps to the locations indicated by the x and y axis positions, either on the same strand (black points) or opposite strands (red points). Green vertical lines in this and subsequent panels delineate the region of high homology surrounding the three glycophorins. **b**) the location of the lead GWAS marker, ABPs, and protein-coding genes in the region. **c**) alignability of the 100-mer at each position of the

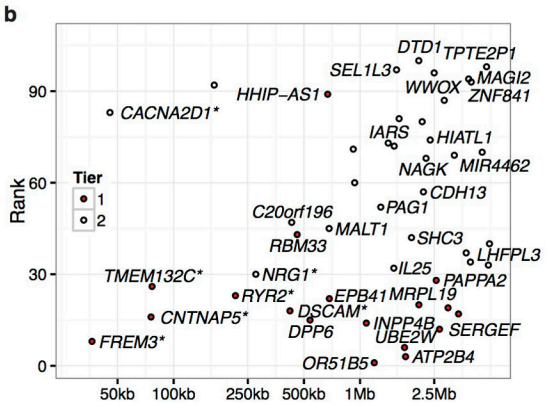
reference, up to two mismatches. Values are taken from the UCSC genome browser mappability track and averaged over 5kb bins. **d**) IMPUTE info in Kenya for variants with frequency at least 5%, averaged over 5kb bins. **e–f**) coverage for samples from YRI and LWK in 1000G Phase 1 carrying esv2662558, carrying esv2668125, or not carrying either deletion, respectively. Coverage for each individual is normalised by the mean coverage for that individual across chromosome 1, and only computed at positions with alignability = 1 for all 100-mers overlapping the position, and for reads with mapping quality at least 20. Values are averaged over 5kb (grey) and 10kb (blue) bins. Three samples with apparently erroneous calls in the 1000G Phase1 genotype release are coloured (NA18519, red; NA19185, yellow; NA19222, green) and assigned to the status indicated by their coverage profile. The bottom row represents a sample of 30 individuals not carrying the deletion selected at random in addition to the two with erroneous genotype calls. Coverage computation was performed using the BAM files available from the 1000G project in October 2014, downloaded from ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data. Four African samples in the Phase 1 release were not assessed because they are not included in this directory.

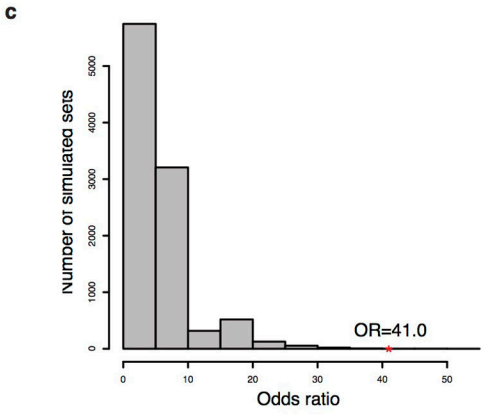



**Extended Data Figure 9.**

Correlation between the genotypes at SNPs of interest within the *GYPE/A/B* locus and reported gene transcription levels in samples from the YRI and LWK HapMap cohorts[17]. *P* values are for a trend test of association where more than one genotype class is present. Only assays targeting the glycophorins, and those with a P-value below 0.05 are shown.

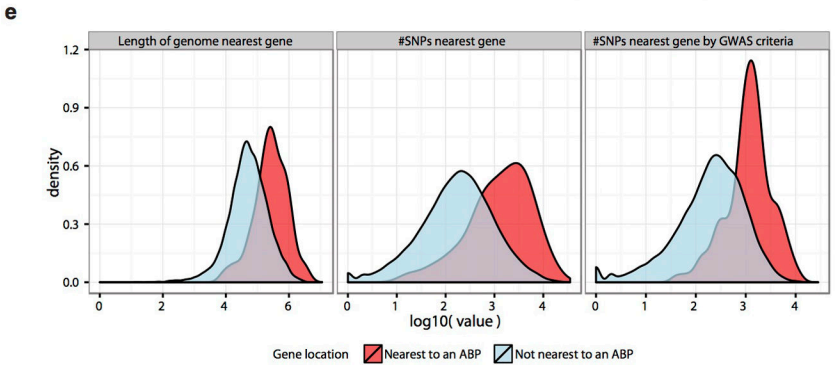



d

| | By gene | By specified distance | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Gene overlap | 50kb | 100kb | 250kb | 750kb | 1Mb | 2.5Mb | 5Mb | 10Mb |
| *Enrichment with ABP haplotypes* | | | | | | | | | |
| Tier 1 | $5x10^{-4}$ | 0.18 | 0.005 | 0.011 | 0.0004 | 0.0025 | 0.0065 | 0.025 | 0.11 |
| Tier 1 (frequency matched†) | $1x10^{-4}$ | 0.18 | 0.007 | 0.009 | 0.0006 | 0.0032 | 0.0063 | 0.024 | 0.12 |
| Tier 1 (excluding known regions††) | $6x10^{-4}$ | 1 | 0.035 | 0.031 | 0.003 | 0.012 | 0.097 | 0.14 | 0.23 |
| Tier 2 | 0.15 | 0.33 | 0.54 | 0.55 | 0.59 | 0.53 | 0.09 | 0.17 | 0.27 |
| *Enrichment with ABP coding SNPS* | | | | | | | | | |
| Tier 1 | 0.18 | 0.06 | 0.19 | 0.29 | 0.051 | 0.08 | 0.14 | 0.38 | 0.98 |
| Tier 2 | 0.61 | 1 | 0.51 | 0.20 | 0.50 | 0.61 | 0.13 | 0.27 | 0.61 |




**Extended Data Figure 10.**

Detail of enrichment analysis. **a)** Red histogram: the empirical distribution of the $\log_{10}$ distance of observed tier 1 loci to the nearest ABP haplotype. Grey histogram: distribution of distances for 10,000 simulated tier 1 sets. **b**) The $\log_{10}$ distance of tier 1 (filled red circles) and tier 2 (empty circles) loci to the nearest ABP, plotted against their rank in $BF_{avg}$ order (stronger signals have lower rank). Loci are annotated with the nearest gene where a gene exists within the association region. Asterisks denote nearest genes that are also the nearest gene to an ABP shared haplotype. **c**) Empirical null distribution of the odds ratio for the enrichment of tier 1 loci in the set of genes closest to an ABP shared haplotype, based on

10,000 simulated SNP sets. The red asterisk and text indicate the odds ratio for the observed tier 1 loci. **d**) Distribution of the proportion of the genome which identifies a given gene as nearest, for genes in or not in the set annotated as nearest an ABP haplotype. Left: distribution of the length of the genome for which the given gene is unambiguously the closest gene. Middle: distribution of the number of SNPs in our study for which the given gene is the closest gene. Right: distribution of the number of SNPs in our study for which the given gene is the nearest gene within a recombination interval of 2.5cM±25kb around the SNP, as used to determine nearest genes to GWAS lead SNPs. **e**) Empirical *P*-values for enrichment of ABP haplotypes and coding SNPs in tier 1 and tier 2 GWAS regions. Second column: P-values for enrichment by gene overlap. Third to tenth column: P-values for enrichment by proximity at different length scales. †Results for simulations using SNPs frequency-matched to GWAS tier 1 loci in 1% frequency bins. ††Results for simulations excluding the regions of *ABO, HBB, ATP2B4, FREM3, INPP4B*, and *HHIP-AS1*.

## Acknowledgements

The MalariaGEN Project is supported by the Wellcome Trust (WT077383/Z/05/Z) and the Bill & Melinda Gates Foundation through the Foundations of the National Institutes of Health (566) as part of the Grand Challenges in Global Health Initiative. The Resource Centre for Genomic Epidemiology of Malaria is supported by the Wellcome Trust (090770/Z/09/Z). This research was supported by the Medical Research Council (G0600718; G0600230), the Wellcome Trust Biomedical ethics Enhancement Award (087285) and Strategic Award (096527). Dominic Kwiatkowski receives support from the Medical Research Council (G19/9). Chris C.A. Spencer was supported by a Wellcome Trust Career Development Fellowship (097364/Z/11/Z). The Wellcome Trust also provides core awards to The Wellcome Trust Centre for Human Genetics (075491/Z/04; 090532/Z/09/Z) and the Wellcome Trust Sanger Institute (077012/Z/05/Z and 098051). The Mali MRTC – BMP group is supported by a contract (N01AI85346) and a cooperative agreement (U19AI065683) from the National Institute of Allergy and Infectious Diseases and by a grant (D43TW001589) from the Fogarty International Centre, National Institutes of Health to University of Maryland and University of Bamako and the Mali-NIAID/NIH ICER at USTTB, Mali. Eric Achidi received partial funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement N° 242095 – EVIMalaR and the Central African Network for Tuberculosis, HIV/AIDS and Malaria (CANTAM) funded by the European and Developing Countries Clinical Trials Partnership (EDCTP). Thomas N Williams is funded by Senior Fellowships from the Wellcome Trust (076934/Z/05/Z and 091758/Z/10/Z) and through the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement N° 242095 – EVIMalaR. The KEMRI-Wellcome Trust Programme is funded through core support from the Wellcome Trust. Carolyne Ndila is supported through a strategic award to the KEMRI-Wellcome Trust Programme by the Wellcome Trust (084538). Tanzania/KCMC/JMP received funding from MRC grant number (G9901439). We acknowledge the work of Belco Poudiougou and Amadou Niangaly for their help in collecting clinical data and biological samples for the Bamako study. We thank Luke Jostins and Matti Pirinen for advice on statistical analyses.

## References

1. MalariaGEN. Reappraisal of known malaria resistance loci in a large multi-centre study. Nature genetics. 2014
2. Timmann C, et al. Genome-wide association study indicates two novel resistance loci for severe malaria. Nature. 2012
3. Band G, et al. Imputation-based meta-analysis of severe malaria in three African populations. PLoS genetics. 2013; 9:e1003509. [PubMed: 23717212]
4. Leffler EM, et al. Multiple instances of ancient balancing selection shared between humans and chimpanzees. Science. 2013; 339:1578–1582. [PubMed: 23413192]
5. Fry AE, et al. Common variation in the ABO glycosyltransferase is associated with susceptibility to severe Plasmodium falciparum malaria. Human molecular genetics. 2008; 17:567–576. [PubMed: 18003641]
6. Abecasis GR, et al. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012; 491:56–65. [PubMed: 23128226]

7. Teo YY, Small KS, Kwiatkowski DP. Methodological challenges of genome-wide association analysis in Africa. Nature reviews. Genetics. 2010; 11:149–160.
8. Manske M, et al. Analysis of Plasmodium falciparum diversity in natural infections by deep sequencing. Nature. 2012; 487:375–379. [PubMed: 22722859]
9. Jallow M, et al. Genome-wide and fine-resolution association analysis of malaria in West Africa. Nature genetics. 2009; 41:657–665. [PubMed: 19465909]
10. Blumenfeld OO, Huang CH. Molecular genetics of the glycophorin gene family, the antigens for MNSs blood groups: multiple gene rearrangements and modulation of splice site usage result in extensive diversification. Human mutation. 1995; 6:199–209. [PubMed: 8535438]
11. Sim BK, Chitnis CE, Wasniowska K, Hadley TJ, Miller LH. Receptor and ligand domains for invasion of erythrocytes by Plasmodium falciparum. Science. 1994; 264:1941–1944. [PubMed: 8009226]
12. Mayer DC, et al. Glycophorin B is the erythrocyte receptor of Plasmodium falciparum erythrocyte-binding ligand, EBL-1. Proceedings of the National Academy of Sciences of the United States of America. 2009; 106:5348–5352. [PubMed: 19279206]
13. Baum J, Ward RH, Conway DJ. Natural selection on the erythrocyte surface. Molecular biology and evolution. 2002; 19:223–229. [PubMed: 11861881]
14. Ko WY, et al. Effects of natural selection and gene conversion on the evolution of human glycophorins coding for MNS blood polymorphisms in malaria-endemic African populations. American journal of human genetics. 2011; 88:741–754. [PubMed: 21664997]
15. Tarazona-Santos E, et al. Population genetics of GYPB and association study between GYPB*S/s polymorphism and susceptibility to P. falciparum infection in the Brazilian Amazon. PloS one. 2011; 6:e16123. [PubMed: 21283638]
16. Wang HY, Tang H, Shen CK, Wu CI. Rapidly evolving genes in human. I. The glycophorins and their possible role in evading malaria parasites. Molecular biology and evolution. 2003; 20:1795–1804. [PubMed: 12949139]
17. Stranger BE, et al. Patterns of cis regulatory variation in diverse human populations. PLoS genetics. 2012; 8:e1002639. [PubMed: 22532805]
18. Gurdasani D, et al. The African Genome Variation Project shapes medical genetics in Africa. Nature. 2014
19. Lanzillotti R, Coetzer TL. The 10 kDa domain of human erythrocyte protein 4.1 binds the Plasmodium falciparum EBA-181 protein. Malaria journal. 2006; 5:100. [PubMed: 17087826]
20. Segurel L, et al. The ABO blood group is a trans-species polymorphism in primates. Proceedings of the National Academy of Sciences of the United States of America. 2012; 109:18493–18498. [PubMed: 23091028]
21. Su Z, Cardin N. Donnelly P, Marchini J. the Wellcome Trust Case Control C. A Bayesian Method for Detecting and Characterizing Allelic Heterogeneity and Boosting Signals in Genome-Wide Association Studies. 2009:430–450.
22. Lee PH, O'Dushlaine C, Thomas B, Purcell SM. INRICH: interval-based enrichment analysis for genome-wide association studies. Bioinformatics. 28:1797–1799. [PubMed: 22513993]
23. Karlsson EK, Kwiatkowski DP, Sabeti PC. Natural selection and infectious disease in human populations. Nature reviews. Genetics. 2014; 15:379–393.
24. Otto TD, et al. Genome sequencing of chimpanzee malaria parasites reveals possible pathways of adaptation to human hosts. Nature communications. 2014; 5:4754.
25. Liu W, et al. Origin of the human malaria parasite Plasmodium falciparum in gorillas. Nature. 2010; 467:420–425. [PubMed: 20864995]
26. Prugnolle F, et al. A fresh look at the origin of Plasmodium falciparum, the most malignant malaria agent. PLoS pathogens. 2011; 7:e1001283. [PubMed: 21383971]
27. Rich SM, et al. The origin of malignant malaria. Proceedings of the National Academy of Sciences of the United States of America. 2009; 106:14902–14907. [PubMed: 19666593]
28. Joy DA, et al. Early origin and recent expansion of Plasmodium falciparum. Science. 2003; 300:318–321. [PubMed: 12690197]

29. Gilbert SC, et al. Association of malaria parasite population structure, HLA, and immunological antagonism. Science. 1998; 279:1173–1177. [PubMed: 9469800]
30. Binks RH, et al. Population genetic analysis of the Plasmodium falciparum erythrocyte binding antigen-175 (eba-175) gene. Molecular and biochemical parasitology. 2001; 114:63–70. [PubMed: 11356514]
31. Severe falciparum malaria. World Health Organization, Communicable Diseases Cluster. Transactions of the Royal Society of Tropical Medicine and Hygiene. 2000; 94(Suppl 1):S1–S90. [PubMed: 11103309]
32. Teo YY. Genotype calling for the Illumina platform. Methods Mol Biol. 2012; 850:525–538. [PubMed: 22307718]
33. Giannoulatou E, Yau C, Colella S, Ragoussis J, Holmes CC. GenoSNP: a variational Bayes within-sample SNP genotyping algorithm that does not require a reference population. Bioinformatics. 2008; 24:2209–2214. [PubMed: 18653518]
34. Kent WJ. BLAT--the BLAST-like alignment tool. Genome research. 2002; 12:656–664. Article published online before March 2002. [PubMed: 11932250]
35. Bellenguez C, et al. A robust clustering algorithm for identifying problematic samples in genome-wide association studies. Bioinformatics. 2012; 28:134–135. [PubMed: 22057162]
36. Astle W, Balding DJ. Population Structure and Cryptic Relatedness in Genetic Association Studies. Stat Sci. 2009; 24:451–471.
37. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature. 2007; 447:661–678. [PubMed: 17554300]
38. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2015.
39. Ross MT, et al. The DNA sequence of the human X chromosome. Nature. 2005; 434:325–337. [PubMed: 15772651]
40. Delaneau O, Zagury JF, Marchini J. Improved whole-chromosome phasing for disease and population genetic studies. Nature methods. 2013; 10:5–6. [PubMed: 23269371]
41. Howie B, Fuchsberger C, Stephens M, Marchini J, Abecasis GR. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nature genetics. 2012; 44:955–959. [PubMed: 22820512]
42. Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS genetics. 2009; 5:e1000529. [PubMed: 19543373]
43. Wakefield J. Bayes factors for genome-wide association studies: comparison with P-values. Genetic epidemiology. 2009; 33:79–86. [PubMed: 18642345]
44. Carrel L, Cottle AA, Goglin KC, Willard HF. A first-generation X-inactivation profile of the human X chromosome. Proceedings of the National Academy of Sciences of the United States of America. 1999; 96:14440–14444. [PubMed: 10588724]
45. Pirinen M, Donnelly P, Spencer CCA. Efficient Computation with a Linear Mixed Model on Large-Scale Data Sets with Applications to Genetic Studies. Ann Appl Stat. 2013; 7:369–390.
46. Consortium GT. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science. 2015; 348:648–660. [PubMed: 25954001]
47. Welter D, et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic acids research. 2014; 42:D1001–D1006. [PubMed: 24316577]
48. Karolchik D, et al. The UCSC Table Browser data retrieval tool. Nucleic acids research. 2004; 32:D493–D496. [PubMed: 14681465]
49. Manjurano A, et al. USP38, FREM3, SDC1, DDC, and LOC727982 Gene Polymorphisms and Differential Susceptibility to Severe Malaria in Tanzania. The Journal of infectious diseases. 2015
50. Wellcome Trust Case Control, C. Bayesian refinement of association signals for 14 loci in 3 common diseases. Nature genetics. 2012; 44:1294–1301. [PubMed: 23104008]
51. Hillier LW, et al. Generation and annotation of the DNA sequences of human chromosomes 2 and 4. Nature. 2005; 434:724–731. [PubMed: 15815621]

52. Zeller T, et al. Genetics and beyond--the transcriptome of human monocytes and disease susceptibility. PloS one. 2010; 5:e10693. [PubMed: 20502693]

## Author contributions

**Writing group**: Gavin Band1, Dominic P. Kwiatkowski1,2, Kirk A. Rockett1,2, Chris C.A. Spencer1

**Data analysis**: Gavin Band1, Geraldine M. Clarke1, Katja Kivinen2, Quang Si Le1, Ellen M. Leffler1, Dominic P. Kwiatkowski1,2, Kirk A. Rockett1,2, Chris C.A. Spencer1

**Project management**: Kirk A. Rockett1,2, Chris C.A. Spencer1, Victoria Cornelius1, David J. Conway3,24, Thomas N. Williams15,16, Terrie Taylor23, Dominic P. Kwiatkowski1,2

**Study site lead investigators**: David J. Conway3,24, Kalifa A. Bojang3, Ogobara Doumbo5, Mahamadou A. Thera5, David Modiano6, Sodiomon B. Sirima7, Michael D. Wilson9, Kwadwo A. Koram9, Tsiri Agbenyega10,22, Eric Achidi14,20, Thomas N. Williams15,16, Kevin Marsh15, Hugh Reyburn17,18, Chris Drakeley17,18, Eleanor Riley18, Terrie Taylor23, Malcolm Molyneux19

**Clinical data and sample collection**: Muminatou Jallow3,4, Kalifa A. Bojang3, David J. Conway3,24, Margaret Pinder3, Ogobara Doumbo5, Mahamadou A. Thera5, Ousmane B. Toure5, Salimata Konate5, Sibiri Sissoko5, Edith C. Bougouma7, Valentina D. Mangano6, David Modiano6, Sodiomon B. Sirima7, Lucas N. Amenga-Etego8, Anita K. Ghansah9, Abraham V. O. Hodgson8, Kwadwo A. Koram9, Michael D. Wilson9, Tsiri Agbenyega10,22, Daniel Ansong10,22, Anthony Enimil10, Jennifer Evans11,12, Eric Achidi14,20, Tobias O. Apinjoh13, Alexander Macharia15, Kevin Marsh15, Carolyne M. Ndila15, Charles Newton15, Norbert Peshu15, Sophie Uyoga15, Thomas N. Williams15,16, Chris Drakeley17,18, Alphaxard Manjurano17,18, Hugh Reyburn17,18, Eleanor Riley18, David Kachala19, Malcolm Molyneux19, Vysaul Nyirongo19, Terrie Taylor23

**Sample processing, genotyping, data management and project coordination**: Kirk A. Rockett1,2, Katja Kivinen2, Daniel Mead2, Eleanor Drury2, Sarah Auburn1, Susana G. Campino2, Bronwyn MacInnis2, Jim Stalker2, Emma Gray2, Christina Hubbart1, Anna E. Jeffreys1, Kate Rowlands1, Alieu Mendy1, Rachel Craik1, Kathryn Fitzpatrick1, Sile Molloy1, Lee Hart1, Robert Hutton1, Angeliki Kerasidou1,21, Kimberly J. Johnson1, Victoria Cornelius1

### Author affiliations

[1] Wellcome Trust Centre for Human Genetics, University of Oxford, Roosevelt Drive, Oxford OX3 7BN, UK.

[2] The Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1SA, UK.

[3] Medical Research Council Unit, Atlantic Boulevard, Serrekunda, The Gambia.

[4] Royal Victoria Teaching Hospital, Independence Drive, Banjul, The Gambia.

[5] Malaria Research and Training Centre, Faculty of Medicine University of Bamako Bamako Mali.

[6] University of Rome La Sapienza, Italy.

[7] Centre National de Recherche et de Formation sur le Paludisme (CNRFP), Ouagadougou, Burkina Faso.

[8] Navrongo Health Research Centre, Navrongo, Ghana.

[9] Noguchi Memorial Institute for Medical Research, Accra, Ghana.

[10] Komfo Anokye Teaching Hospital, Kumasi, Ghana

[11] Department of Molecular Medicine, Bernhard Nocht Institute for Tropical Medicine, Postfach 30 41 2, D-20324 Hamburg, Germany.

[12] Kumasi Centre for Collaborative Research, Kumasi, Ghana.

[13] Department of Biochemistry & Molecular Biology, University of Buea, Buea, South West Region, Cameroon.

[14] Department of Medical Laboratory Sciences, University of Buea, Buea, South West Region, Cameroon.

[15] KEMRI-Wellcome Trust Research Programme, PO Box 230, Kilifi, Kenya.

[16] Faculty of Medicine, Department of Medicine, Imperial College, Exhibition Road, London SW7 2AZ, UK.

[17] Joint Malaria Programme, Kilimanjaro Christian Medical Centre, PO box 2228, Moshi, Tanzania.

[18] Faculty of Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine, Keppel Street, London, UK.

[19] Malawi-Liverpool-Wellcome Trust Clinical Research Programme, College of Medicine, University of Malawi, PO Box 30096, Blantyre, Malawi.

[20] Weatherall Institute of Molecular Medicine, Oxford University, Oxford, UK.

[21] The Ethox Centre, Nuffield Department of Population Health, University of Oxford, Old Road Campus, Oxford, OX3 7LF.

[22] Kwame Nkrumah University of Science and Technology, Kumasi, Ghana.

[23] Blantyre Malaria Project, College of Medicine, University of Malawi.

[24] Department of Pathogen Molecular Biology, London School of Hygiene and Tropical Medicine, Keppel Street, London, WC1E 7HT, UK.

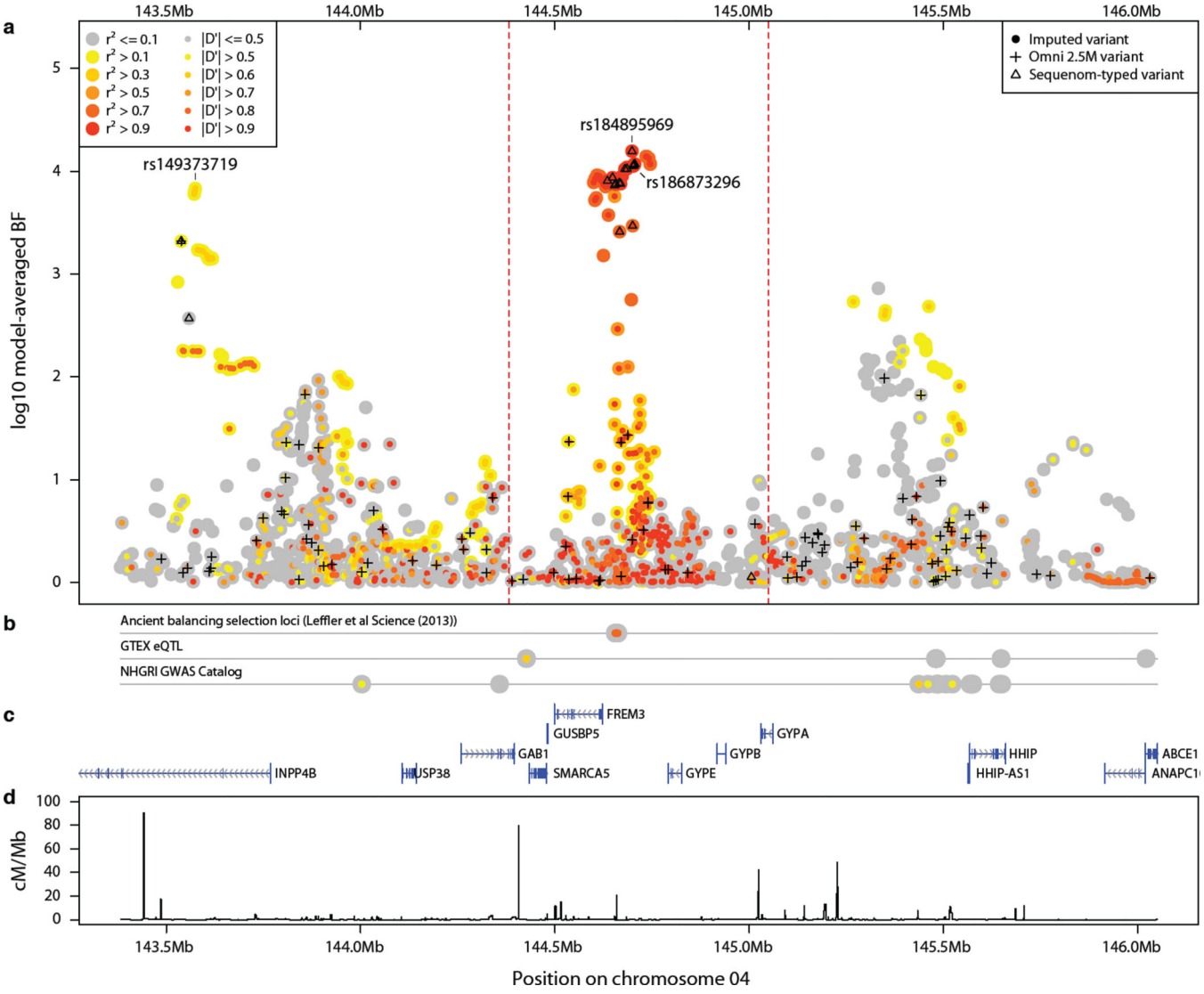



**Figure 1.**

Signal of association with severe malaria across the *FREM3/GYPE* region. **a**) evidence for association ($\log_{10}BF_{avg}$) in the discovery data. Black plusses denote SNPs that were directly typed, and black triangles denote SNPs selected for typing on the Sequenom platform. Dotted red vertical lines indicate a region of 0.25cM±25kb centred at the lead SNP (rs184895969). Coloured circles denote the correlation (outer circles) and |D'| (inner circles) with rs184895969 in controls, computed from imputed haplotypes. **b**) Polymorphisms shared between humans and chimpanzees, eQTLs, and previously reported associations with other phenotypes. **c,d**) Genes in the region and the HapMap combined recombination rate.

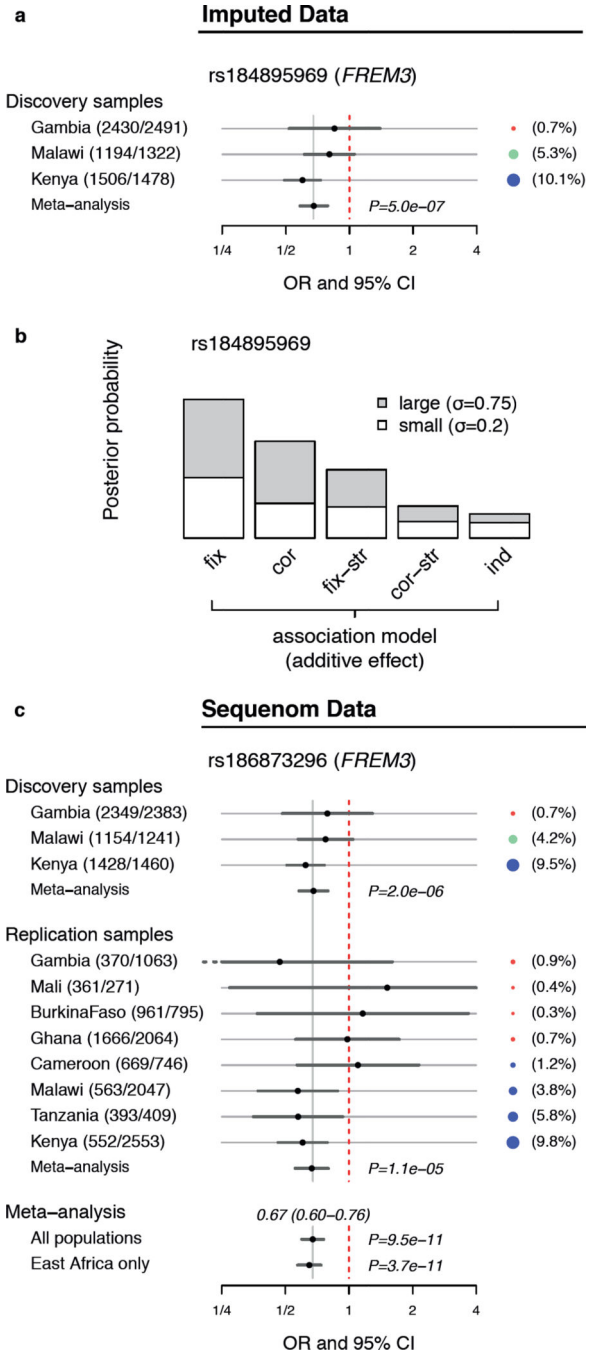



**Figure 2.**

Evidence for association at SNPs in the *FREM3/GYPE* region assuming an additive model of association. **a**) Forest plot showing sample size, estimated odds ratio and 95% confidence interval for the lead imputed SNP in each population and under fixed-effect meta-analysis. The frequency of the protective allele in controls in each population is shown to the right. **b**) The posterior weight on different models of heterogeneity at rs184895969 under the prior used in the GWAS. Model names are described in Methods. **c**) Forest plot for the Sequenom-typed SNP rs186873296 in discovery and replication samples, with fixed-effect

meta-analysis across all populations and across East African populations (here taken as Kenya, Malawi, Tanzania and Cameroon.)

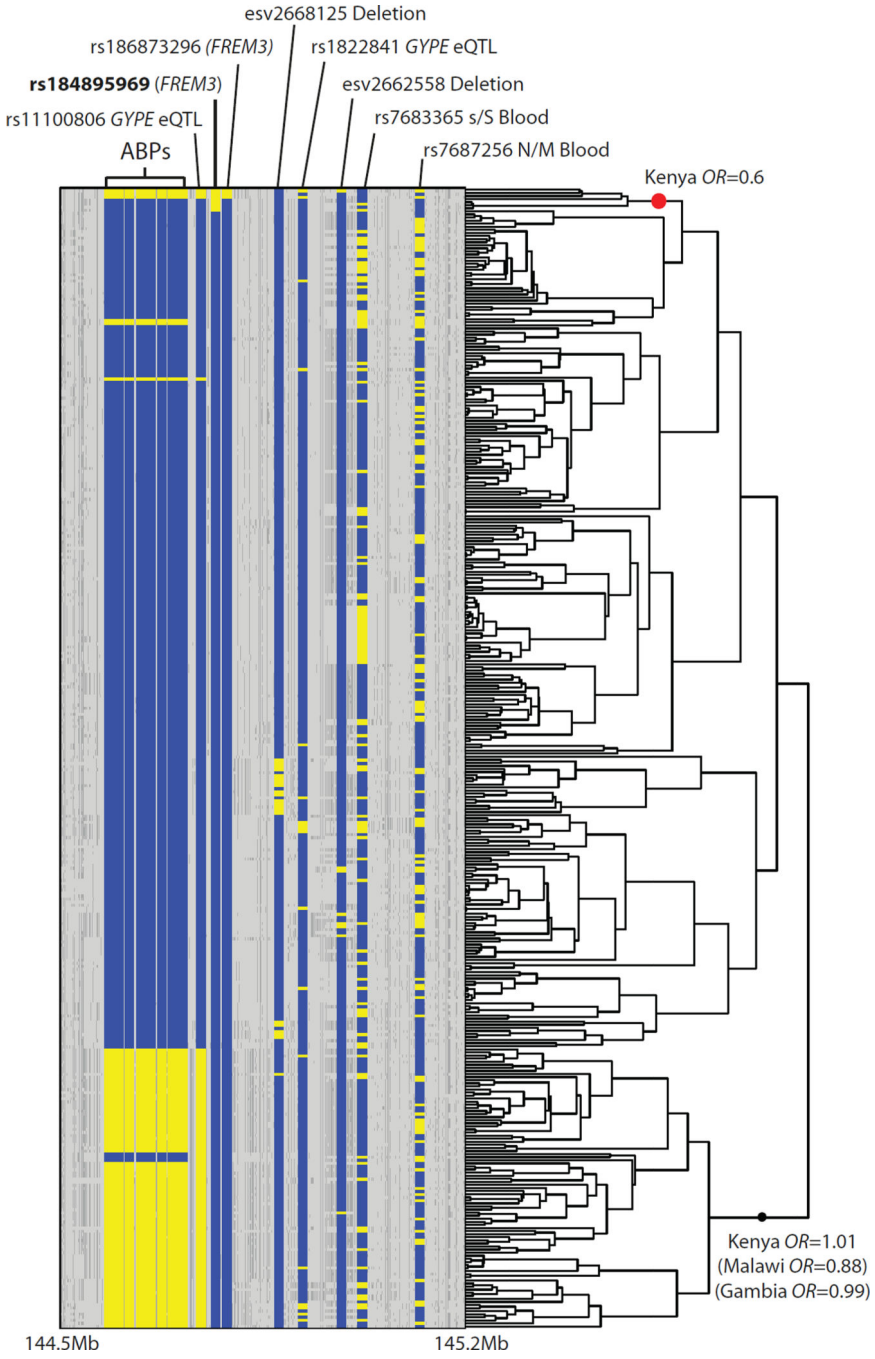



**Figure 3.**

Haplotype analysis across the *FREM3/GYPE* region. Left hand panel shows haplotypes at 7321 polymorphic SNPs between 144.5Mb and 145.2Mb on chromosome 4 in the LWK and YRI samples of the 1000 Genomes reference panel. Key variants (Methods and Supplementary Note 2) are enlarged for clarity and labelled, with reference and non-reference alleles coloured blue and yellow respectively. On the right is the estimated topology of the genealogical tree at rs184895969. Dots indicate the position of the inferred

protective mutation in Kenya and the branch ancestral to the ABPs, and are labelled with the estimated odds ratios (*OR*).